A pseudomolecule-scale genome assembly of the liverwort Marchantia polymorpha

Diop, Seydina I ; Subotic, Oliver ; Giraldo-Fonseca, Alejandro ; Waller, Manuel ; Kirbis, Alexander ; Neubauer, Anna ; Potente, Giacomo ; Murray-Watson, Rachel ; Boskovic, Filip ; Bont, Zoe ; Hock, Zsofia ; Payton, Adam C ; Duijsings, Daniël ; Pirovano, Walter ; Conti, Elena ; Grossniklaus, Ueli ; McDaniel, Stuart F ; Szövényi, Péter

DOI: https://doi.org/10.1111/tpj.14602

DR PETER SZOEVENYI (Orcid ID : 0000-0002-0324-4639)

Article type : Original Article

A pseudomolecule-scale genome assembly of the liverwort Marchantia polymorpha

Seydina Issa Diop[1,2], Oliver Subotic[1,2], Alejandro Giraldo-Fonseca[3], Manuel Waller[1], Alexander Kirbis[1], Anna Neubauer[1], Giacomo Potente[1,2], Rachel Murray-Watson[1], Filip Boskovic[1], Zoe Bont[1], Zsofia Hock[1], Adam C Payton[4], Daniël Duijsings[2], Walter Pirovano[2], Elena Conti[1], Ueli Grossniklaus[3], Stuart F McDaniel[4], Péter Szövényi[1]

[1]Department of Systematic and Evolutionary Botany & Zurich-Basel Plant Science Center, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland

[2]BaseClear B.V.,Sylviusweg 74 ,2333 BE, LEIDEN, The Netherlands

[3]Department of Plant and Microbial Biology & Zurich-Basel Plant Science Center, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland

[4]Department of Biology, University of Florida, 876 Newell Drive, Gainesville, Florida 32611 USA

Author for correspondence: Peter Szovenyi, peter.szoevenyi@uzh.ch, Dept. of Systematic and Evolutionary Botany, University of Zurich, Zollikerstr 107, 8008 Zurich, Switzerland. Phone: +41 (0) 634 84 40

Running head: Pseudomolecule-scale genome assembly of *M. polymorpha*

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/TPJ.14602

Key words: DNA methylation, pseudomolecule, evolution, large-scale genome structure, bryophytes, recombination rate

ABSTRACT

Marchantia polymorpha has recently become a prime model for cellular, evo-devo, synthetic biological, and evolutionary investigations.

We present a pseudomolecule-scale assembly of the *M. polymorpha* genome making comparative genome structure analysis and classical genetic mapping approaches feasible. We anchored 88% of the *M. polymorpha* draft genome to a high-density linkage map resulting in eight pseudomolecules. We found that the overall genome structure of *M. polymorpha* is in some respects different from that of the model moss *Physcomitrella patens*. Specifically, genome collinearity between the two bryophyte genomes and vascular plants is limited suggesting extensive rearrangements since divergence. Furthermore, recombination rates are greatest in the middle of the chromosome arms in *M. polymorpha* like in most vascular plant genomes, which is in contrast to *P. patens* where recombination rates are evenly distributed along the chromosomes. Nevertheless, some other properties of the genome are shared with *P. patens*. As in *P. patens*, DNA methylation in *M. polymorpha* is spread evenly along the chromosomes, which is in stark contrast to *Arabidopsis thaliana*, where DNA methylation is strongly enriched at the centromeres. Nevertheless, DNA methylation and recombination rate are anticorrelated in all three

species. Finally, *M. polymorpha* and *P. patens* centromeres are of similar structure and marked by high abundance of retroelements unlike in vascular plants.
Taken together, the highly contiguous genome assembly we present opens unexplored avenues for *M. polymorpha* research by linking the physical and genetic maps, making novel genomic and genetic analyses, including map-based cloning, feasible.

Introduction

The liverwort *Marchantia polymorpha* is one of the oldest models for studying the biology of plants (Schmidel, 1762; Hedwig, 1783; Bowman *et al.*, 2016), including cellular growth, plant development, speciation, and classical genetics (Burgeff, 1943). Studies of *M. polymorpha* played a crucial role in the discovery of the alternation of sporophyte and gametophyte generations (Schmidel, 1762; Hedwig, 1783; Hofmeister, 1851), and contributed to the early discovery of sex chromosomes and sex determination in plants. Asexual propagules of *M. polymorpha* were used to understand how a dorsiventral body plan can develop from initially apolar cells (Mirbel, 1835; Zimmerman, 1882; Oppenheimer, 1922; Fitting, 1936; Halbsguth & Kohlenbach, 1953; Otto, 1976; Otto & Halbsguth, 1976), and studies of *M. polymorpha* were central in inspiring the debate on the cellular nature of organisms and the origin of new cells (Allen, 1917, 1945; Haupt, 1932; Nakayama *et al.*, 2001; Yamato *et al.*, 2007; Jamilena *et al.*, 2008). Relatives of *M. polymorpha* have been extensively used to study fundamental questions of evolutionary ecology and population biology (Stark *et al.*, 2005; Groen *et al.*, 2010; Stieha *et al.*, 2014; Brzyski *et al.*, 2018). More recently, the phylogenetic position of *M. polymorpha* has made this species critical for understanding the evolution of gene regulatory networks across land plants (Breuninger *et al.*, 2016; Bowman *et al.*, 2017; Honkanen *et al.*, 2018). Phylogenomic data now indicate that

liverworts and mosses are monophyletic, and that the tracheophytes (vascular plants) are sister to the clade of bryophytes including mosses, liverworts and hornworts (Wickett *et al.*, 2014; Cox, 2018; Puttick *et al.*, 2018a).

The recently published genome of *M. polymorpha* (Bowman *et al.*, 2017) revealed some striking differences from the genome of the moss, *P. patens* (Lang *et al.*, 2018). The *M. polymorpha* genome is relatively small (230 Mbp vs the 462 Mbp genome of *P. patens*), with a low repeat content and redundancy, which leads to a significantly reduced gene set especially for regulatory genes compared to *P. patens*. However, the current assembly contains 2,957 unordered scaffolds, which precludes chromosome-scale genome structure comparisons with *P. patens*. For example, in contrast with vascular plant genomes sequenced to date, the *P. patens* genome showed an even distribution of gene, repeat density, and recombination rate along the chromosomes (Lang *et al.*, 2018). It is currently unclear whether this is a unique feature of the *P. patens* genome or it is shared with other bryophyte groups, such as the liverworts. Furthermore, the lack of a highly contiguous *M. polymorpha* assembly hinders classical forward genetics approaches, which could proceed rapidly given that a single cross produces many recombinant haploid progeny (McDaniel *et al.*, 2007; Kamisugi *et al.*, 2008).

Here we used genetic mapping to generate a highly contiguous genome assembly for *M. polymorpha.* We created a mapping population from a cross between the genome isolate and a Swiss isolate and constructed a high-density linkage map using over 4,000 genetic markers obtained by ddRAD-seq (Parchman et al., 2012). We then anchored the genetic map to the v1.3 assembly (Bowman et al., 2017) arranging scaffolds into eight linkage groups corresponding to the eight autosomes of the *M. polymorpha* genome. We used this assembly to study the genome organization of *M. polymorpha*. At a large-scale, the recombination landscape of *M. polymorpha* was more similar to those of vascular plant genomes than that of *P. patens*. Despite that, some other properties of the genome are shared with *P. patens* which may be ancestral to the group of bryophytes. For instance, the genome-wide pattern of DNA-methylation shows a more even distribution along the chromosome arms than in *Arabidopsis thaliana*, a feature shared with *P. patens*, but as in *A. thaliana* recombination frequency anticorrelates with DNA-methylation levels. Similarly, gene and repeat density are more evenly distributed along the chromosome arms in *M. polymorpha* and *P. patens* than in *A. thaliana*. We also found that *M. polymorpha* and *P. patens* centromeres are of similar structure and marked by high abundance of retroelements unlike in

vascular plants. Nevertheless, genomic collinearity between the two bryophyte species and vascular plant genomes was very limited.

Results

ddRad-tag generation and linkage map construction

The sequencing runs generated 151,226,066 and 146,065,074 raw reads, respectively. After demultiplexing, quality filtering, and trimming, we obtained 292235916 reads (approx. 98% of the raw reads). We retained 3,792,663 (Tak1) and 4,441,547 (IrchF) reads for the parental lines. From the segregating population, we had to exclude or discard 7 samples due to low sequencing coverage or because their library preparation failed in two consecutive attempts.

Using the filtered and trimmed sequence data and the refmap.pl script, we obtained 6601 loci. After dropping four individuals by merging genetic clones, discarding 12 markers with significant segregation distortion, and removing markers with more than 20% missing data, the mapping data contained 4,536 markers (ddRAD-tags) and 79 individuals plus the two parental lines.

In our linkage map, the 4,536 markers were assigned to eight linkage groups (LG) (Fig. 1, a), which corresponds to the number of autosomes in *M. polymorpha* based on cytological observations (Ono, 1976; Bischler, 1986). The number of LGs was robust against the LOD p-values and the missing data thresholds used. Furthermore, a heat plot of LOD scores and pairwise recombination rate estimates between markers suggests that our map is solid and shows no obvious problems (Fig. 1, b). The eight linkage groups were 81.06 centimorgan (cM), 83.60 cM, 69.76 cM, 102.65 cM, 101.51 cM, 111.49 cM, 76.01 cM, and 86.15 cM with an average of 89.03 cM and with a total length of 712.23 cM (Table 1). Marker distances showed little variation across chromosomes with an overall average of 1.76 cM and a minimum and maximum value of 1.58 and 1.90 cM. 88% of the draft assembly (nucleotides) was assigned to one of the eight linkage groups, consisting of 381 scaffolds with a total length of 199,989,030 Mb. Conversely, 12% of the bases of the draft assembly remained unassigned.

Anchoring the linkage map to the assembly

After linkage map construction, we anchored and ordered scaffolds of the draft assembly on the linkage map. We discarded markers with ambiguous and low-quality mapping to the draft genome and used 4,234 markers for anchoring. We were able to anchor about 88% of the total bases of the linkage map. 12% of the total base space that were part of the eight linkage groups remained unplaced in the final map. We could lift over all gene models (19,287) of the *M. polymorpha* draft genome (Bowman *et al.*, 2017) to our assembly of which 94.6% (18,247) mapped to the eight linkage groups. Nonparametric correlation statistics (Spearman's Rho) between physical position of the markers and their genetic map position was greater than 0.97 for each of the eight LGs (Table 1), indicating a good fit between the two map types and success of the anchoring process (Supplementary Information Fig. S1). Nevertheless, the anchoring also revealed some conflicts between physical and genetic positions of markers, which are either due to mis-assemblies or errors in the linkage map. For instance, physical and genetic position of markers were in conflict in the terminal regions of linkage groups LG1, LG2, and LG3 (Supplementary Information Fig. S1).

When we filled between-scaffold gaps using an arbitrary number of 100 Ns, the final map had a total length of 198,443,496 Mb (excluding gaps), in which scaffolds of the draft genome are arranged into eight pseudomolecules. Size of the pseudomolecules varied between 20.5 and 29.5 Mb, with an average of 24.80 Mb. The two Y chromosome scaffolds and another 2,596 scaffolds remained unplaced on the map. When we estimated the size of inter-scaffold gaps using the genetic map, the total length of the map increased to 208,823,686 Mb and pseudomolecule sizes were between 30.69 and 21.42 Mb, with an average of 26.10 Mb. That is, including the estimated size of inter-scaffold gaps in the map added about 10 Mb to the total length of the assembly.

Telomeric tandem repeats

We found that the conserved plant telomeric repeat (TTTAGGG)n, was present on only one end of a single linkage group (LG2: $(TTTAGGG)_{35}$). We did not detect less common centromeric repeats described for green plants such as, the one found in *Chlamydomonas* [TTTTTAGGG] (Fulneckova *et al.*, 2016), *Cestrum* [TTTTTTAGGG] (Sykorova *et al.*, 2003), *Allium* [CTCGGTTATGGG] (Fajkus *et al.*, 2016) or in other plants having the human type of repeat [TTAGGG](Supplementary Information Table S1). When we plotted the distribution of the

number of repeat units per tandem array, we found that the longest arrays were located towards the end of the LGs (Supplementary Information Fig S2). Nevertheless, these longest tandem arrays were rarely terminally located and even when assessing the position of shorter tandem arrays (assuming that assembly of long tandem arrays is rarely complete), they did not show specific affinity to the telomeres. Also, there was not a specific repeat sequence that would have been associated with all telomeric regions in general. That is, tandem repeats with a repeat unit length of 5-15 bp do not seem to be specifically associated with the telomeric regions of the linkage groups assembled in *M. polymorpha*.

Centromeric repeats

Typical plant centromeric repeats are composed of relatively long repeat units (50-500bp) that are organized into long tandem arrays (hundreds to thousands of copies per array) (Melters *et al.*, 2013; Oliveira & Torres, 2018; Hartley & O'Neill, 2019). We found that both distribution of repeat unit length and the number of repeat units per tandem array was highly skewed to the left (highly abundant short tandem repeats representing minisatellites and microsatellites) with few repeat units longer than 18 bps and very few arrays containing more than 20 tandem repeats per array (Supporting Information Fig S3). Therefore, long tandem repeats with high copy number are not present in the genomic sequence. We then assessed whether density of potential centromeric tandem arrays ([a] repeat unit length ≥10bps and number of repeat units per tandem array≥10, [b] repeat unit length ≥10bps and number of repeat units per tandem array≥30; see methods) is greater around the putative centromeric regions than elsewhere (see section on Marey maps and identification of centromeric regions below). The two tandem repeat array classes (see [a] and [b] above) showed increased frequency around the putative centromeric regions (see below) only on one or two LGs (Supplementary Information Fig S4). Finally, we also searched a candidate centromeric repeat against the assembled genome sequence reported by (Melters *et al.*, 2013). We found that the putative centromeric repeat sequence mapped to the putative centromeric regions in only four out of eight LGs (LG2-3,LG7-8; Supplementary Information Fig S4). Furthermore, this putative centromeric repeat was not present in high copy numbers typical for flowering plants.

Marey maps, recombination rate variation

Our Marey maps indicated that recombination rate variation along each of the eight assembled pseudomolecules of *M. polymorpha* deviated significantly from zero ($p \leq 0.0001$ for each of them, Supplementary Information Table S2). In general, all eight pseudomolecules showed one or more recombination active and inactive region (Fig. 2). We used these regions to identify putative centromeric regions, which are expected to be devoid of recombination (Copenhaver *et al.*, 1999; Vincenten *et al.*, 2015; Nambiar & Smith, 2016) (Fig. 2). The number of active and inactive regions were variable across the chromosomes. In particular, for LG1, LG2, LG4, LG5, and LG6 our analyses suggested one major extensive region with a very low recombination rate. In LG3, LG7, and LG8, there are at least two regions with similarly low recombination rates. Recombinationally inactive regions are either located close to the middle or close to one end of the pseudomolecules. Therefore, the position of centromeres could be readily defined for LG1, LG2, LG3-LG6, but not for LG7 and LG8, due to the presence of at least two regions with highly suppressed recombination. For these latter LGs, the regions with the lowest and most extensive recombination suppression was identified as putative centromeres (Fig. 2). In LG7, the putative centromeric region was also supported by a peak in DNA methylation. Nevertheless, localization of the centromere in LG8 using recombination rate variation remained ambiguous even after taking the distribution of DNA-methylation into account. Recombination rate variation was also significantly different from zero along all 23 pseudomolecules of *P. patens* ($p \leq 0.0001$ for each of the chromosomes, Supplementary Information Table S3 and Fig. S5). We excluded chromosomes 6, 13, 25, and 27 from the analysis as explained in the methods section.

Distribution of transposable elements and putative centromeres

As described above, we first identified putative centromeric regions using recombination rate variation along the eight linkage groups. We then tested whether maximum density of transposons was associated with these putative centromeres, a pattern reported for the *P. patens* genome (Lang *et al.*, 2018). We further argued that if a relationship between transposon density and centromeres exists, it will help us to refine the localization of centromeres. Peaks in the density of all repeat elements along the eight linkage groups did not coincide with the putative position of centromeres (Supporting Information Fig S6 and Fig. 3). For the majority of the chromosomes, retrotransposon

density peaks coincided with the position of putative centromeres identified based on DNA-methylation status and Marey maps (Supporting Information Fig. S6). The frequency of all retrotransposons was higher for all putative centromeric regions compared to their surroundings. Nevertheless, retrotransposon density was not necessarily the highest in the centromeric region. We found that maximum density of non-LTR retrotransposons, especially those of LINE elements, coincided with the position of the putative centromeric regions for most of the linkage groups (LG2-5 and LG7-8). In contrast, the distribution of unclassified and unknown elements as well as DNA transposons preferentially avoided putative centromeric regions. Moreover, maximum density of all DNA transposons, LTR Copia and LTR Gypsy retrotransposons did not coincide with the position of putative centromeres. Distribution of LINE elements helped to clarify the position of centromeres on LG7 and LG8 which was ambiguous only using information on recombination rates. Nevertheless, on LG1 and LG6 the maximum density of LINE elements did not coincide with the putative centromeric region but with a segment of the chromosome representing another low recombination valley. Therefore, uncertainty remains concerning the position of the centromeres on LG1 and LG6. We took this uncertainty into account (see next paragraph) and carried out all further analyses using centromere positions obtained by recombination rate variation and by the maximum density of LINE elements. We note, that changing centromere positions on LG1 and LG6 did not qualitatively change the outcome of the tests, nevertheless, it did influence the actual values of the statistics. Therefore, test statistics presented in Table 2 are based on centromeric positions obtained using our Marey map data.

Recombination rate variation

We then tested whether recombination rates and selected genomic features show significant variation along the chromosome arms at a large scale (among the four quantiles of the chromosome arms). The nested ANOVA revealed that recombination rate shows an overall significant variation among the four quartiles of the chromosome arms in *M. polymorpha* (Table 2). This implies that recombination rate is not evenly distributed across the chromosome arms at a large scale, and it is usually higher in the middle of the arms and lower close to the centromeres and telomeres. In contrast, most genomic features investigated did not show significant variation among the four quartiles of the chromosome arms. In particular, the number of gene models, the

proportion of bases covered by gene models, the proportion of bases covered by repeats in 500kb windows, and DNA-methylation in all three contexts (CG, CHG, and CHH, with H being any nucleotide but G) were not significantly different among the four regions of the chromosome arms. Only GC content in 500kb windows turned out to be significantly different among the four regions, showing greater values in the middle than in the pericentromeric and telomeric regions.

In contrast, recombination rates were not significantly different among the four regions of the chromosome arms in *P. patens*, implying that recombination rates are more evenly distributed in *P. patens* than in *M. polymorpha* at a large scale (Table 2). Nevertheless, like in the *M. polymorpha* genome, most of the genomic features showed no significant variation among the four quartiles of the chromosome arms either. We found that GC content, proportion of bases covered by gene models and by repeats in 500kb windows, and DNA-methylation in all three contexts were not significantly different among the four regions of the chromosome arms. Only the number of gene models in 500kb windows showed a significant variation across the four quartiles but this difference was due to the drop of gene model density at the telomeric regions of the chromosomes.

Correlation between recombination rate and genomic features

In *M. polymorpha*, our analysis indicated that GC content was not significantly correlated with the recombination rate along the chromosomes, even when using the least smoothed recombination rate estimates (Table 3). In contrast, we found that gene density measured both as number of genic features and proportion of bases covered by genes in 500kb windows was significantly positively correlated with recombination rate estimates, when using recombination rate estimates obtained with the least smoothing. We also observed that the proportion of bases occupied by repeat features per window was significantly negatively correlated with recombination rate estimates, regardless of the degree of smoothing used.

In contrast to *M. polymorpha*, GC content was significantly and positively correlated with recombination rate across the chromosomes in *P. patens* (Table 3). The number of gene models

per window, and the proportion of bases covered by gene models, were both significantly positively correlated with the recombination rate, similar to *M. polymorpha*. Conversely, the proportion of bases covered by repeats in the genome was significantly negatively correlated with recombination rates. All correlations were robust against the smoothing used to estimate recombination rates.

Some genomic features were also intercorrelated (Table 3). In particular, we found that the proportion of bases covered by gene and repeat features were strongly negatively correlated in both species. In contrast, GC content and number of gene features per window were significantly positively correlated in both species.

DNA methylation and recombination rate

DNA methylation in 500kb nonoverlapping windows showed considerable variability along the eight linkage groups of *M. polymorpha,* with the CG context showing the most variation (Fig. 2). Peaks in cytosine methylation overlapped with the location of the putative centromeres for most but not for all eight LGs (Fig. 2). Nevertheless, variability in DNA methylation along the eight LGs was less pronounced than in *A. thaliana* (Supporting Information Fig. S7). This was also true for the *P. patens* genome (Supporting Information Fig. S5). However, DNA methylation was significantly negatively correlated with recombination rate along both the *P. patens* and *M. polymorpha* genomes. This correlation was independent of the span value or the methylation context used (Table 4).

Gene body methylated genes

We found that using the thresholds of at least one or ten methylated CpG position per gene body, 3968 and 303 out of the 16677 investigated genes showed gene body methylation, respectively. Both correlation analysis and Wilcox-tests indicated that gene body methylation tends to decrease expression specificity and gene body methylated genes are more broadly expressed than genes with no gene body methylation ($Tau_{[nonmethylated/methylated,\ median]}$=0.7794, $Tau_{[nonmethylated/methylated,\ median]}$=0.7310, $p_{Wilcox\text{-}test}<10^{-10}$; Spearman`s Rho=-0.0523, $p<10^{-6}$). That is, overall increasing gene

body methylation makes genes more broadly expressed but this effect is weak. We further found that number of methylated positions per gene body was positively correlated with gene length, number of exons but negatively correlated with GC content (Spearman`s $Rho_{gene\ length}$=0.2287, $p<10^{-10}$; $Rho_{number\ of\ exons}$=0.2001, $p<10^{-10}$; Spearman`s $Rho_{GCcontent}$=-0.1605, $p<10^{-10}$). Therefore, gene body methylated genes are longer, have more exons and a lower GC content than nonmethylated genes which conclusion was also confirmed when treating gene body methylation as a discrete character (Gene $length_{[nonmethylated/methylated;\ median]}$= 2730/4242, $p_{Wilcox\text{-}test}<10^{-10}$; Exon $number_{[nonmethylated/methylated;\ median]}$= 3/5, $p_{Wilcox\text{-}test}<10^{-10}$; GC $content_{[nonmethylated/methylated;\ median]}$= 0.4729/0.4570, $p_{Wilcox\text{-}test}<10^{-10}$). In contrast, average expression level of gene body methylated and nonmethylated genes did not differ significantly (average $expression_{[nonmethylated/methylated;\ median]}$= 5.7940/6.2952; $p_{Wilcox\text{-}test}$=0.5645) and there was no significant correlation between level of gene body methylation and average level of gene expression (Spearman`s $Rho_{avegare\ expression}$= -0.0073, p=0.3400).

Limited genome-wide collinearity between bryophytes and across bryophytes and vascular plants

Overall, our analysis showed limited collinearity across the two bryophyte and vascular plant genomes investigated. We found no conserved collinear segment across all embryophytes regardless of the number of anchor points used and the number of collinear segments was always greater within vascular plants than between vascular plants and bryophytes (Supporting Information Table S4,S5,S6). Furthermore, collinearity was more limited between the two bryophyte genomes than between each of the bryophyte and vascular plant genomes for each parameter combinations investigated (Supplementary Information Table S4-6).

In particular, requiring five anchor points to be present per collinear block we found one collinear block between *M. polymorpha* and *Theobroma cacao* and one between *P. patens* and *P. trichocarpa*. Nevertheless, no conserved blocks were discovered between *M. polymorpha* and *P. patens*. As expected, there were many more collinear blocks among dicots but less between dicots and monocots (Supporting Information Table S4). Results did not change qualitatively when using a minimum of four anchor points (Supporting Information Table S5).

However, we obtained qualitatively different results when using a minimum of three anchor points in the analysis (Supplementary table S6). We found 97 and 117 collinear blocks between one or more vascular plants and either *M. polymorpha* or *P. patens*, respectively. The majority of these was collinear between one of the bryophyte and one vascular plant species but the second most abundant class included collinear blocks shared with two or four vascular plants. Only one segment was collinear across nine vascular plant species in *M. polymorpha*. In contrast, collinear regions between *P. patens* and vascular plants were shared at most among eight species. The collinear segments conserved across nine (in *M. polymorpha*) or eight (in *P. patens*) species included both dicots and monocots and were spread out over all chromosomes of *M. polymorpha* or *P. patens* (Fig. S8). Importantly, none of these conserved collinear regions were shared between *P. patens* and *M. polymorpha*.

We also assessed the GO enrichment of genes located in the blocks collinear between one of the bryophyte genomes and at least another four vascular plant species genomes. We found that in *M. polymorpha* genes in these blocks were enriched among others for aromatic product metabolism, nucleic acid metabolism especially RNA modifications, photosynthesis and nitrogen compound synthesis (biological processes ontology) (Supporting Information Fig S9). Many more genes were located in these collinear blocks in *P. patens*. Furthermore, genes in these regions were enriched for various GO terms that did not occur in the set collinear between *M. polymorpha* and vascular plants and are potentially related to key vascular plant traits. Among others, we found that terms related to stomata, photosynthesis, embryo development, reproduction, anthocyanin and hormone synthesis/metabolism, recombination, and post-embryonic development, DNA-damage handling, circadian rhythm, water homeostasis showed significant enrichment (Supporting Information Fig. S9 B).

We also found 10 collinear regions that were exclusively shared between the two bryophyte species but not with any of the other vascular plant species. Genes of these segments were part of gene families including members of most vascular plant genomes. Therefore, lack of collinearity was not due to the predominance of bryophyte-specific genes in these segments but rather to the lack of collinearity with vascular plant genomes. These 10 segments in total consisted of 327 genes in *M. polymorpha* and 711 genes in *P. patens*. Half of the co-linear regions occurred on LG6 of *M. polymorpha* (Supporting Information Fig. S8). The rest occurred on LG4,LG1 and LG7. A combined GO enrichment analysis of *P. patens* and *M. polymorpha* genes suggested that these

collinear regions were enriched for biological process terms potentially related to the specialized morphology, life cycle, molecular biology and physiology of bryophytes (Supporting Information Fig S9). For instance, we found that terms related to secondary metabolite production, DNA repair, response to UV and far-red light, ion transport, carbohydrate metabolism and various other metabolic processes were strongly enriched in these regions. Furthermore, genes involved in regulating key developmental processes such as axis formation, meristem maintenance were also enriched.

Discussion

Linkage groups and the karyogram of M. polymorpha

M. polymorpha is a haploid dioecious plant with chromosomal sex determination system. It has eight autosomes and one sex chromosome (either a female-determining U or male-determining V chromosome), the segregation of which during meiosis determines sex (Yamato *et al.*, 2007; Jamilena *et al.*, 2008; Bowman *et al.*, 2016). The karyotype of *M. polymorpha*, like multiple species in the genus *Marchantia* (Haupt, 1932; Bischler, 1986), consists of five or six metacentric and two or three submetacentric autosomes, depending on the thresholds used to classify these two types of chromosomes (Ono, 1976; Bischler, 1986). The *M. polymorpha* sex chromosome that are both likely metacentric (Yamato *et al.*, 2007), but because a genetic map cannot be used to order markers on the sex chromosomes (these chromosomes do not recombine), here we focus on the autosomes.

Quality of the anchored linkage map

The recently published genome of *M. polymorpha* consisted of nearly 3,000 scaffolds (Bowman *et al.*, 2017). Here we provide a linkage map composed of eight linkage groups corresponding to the eight autosomes of the genome. Combining information on the size of the pseudomolecules with

the physical distribution of genomic features, especially with the position of the centromeres, made it possible to establish a putative correspondence between the pseudomolecules and the actual *M. polymorpha* chromosomes (Fig. 3). The relative size of our pseudomolecules fits well with previous cytogenetic observations; nevertheless, proper cytogenetic assignment of the pseudomolecules to chromosomes remains to be performed (Ono, 1976; Bischler, 1986). Having established a correspondence between karyotype and assembly will make it possible to easily locate and visualize regions of interest and to connect observations at the genome and the karyotype level.

Our linkage map compares well with maps from other species, especially in terms of the number of markers and their density. In particular, the *M. polymorpha* map is based on over 4,500 genetic markers, which is slightly more than what was used to build the linkage map of the moss *P. patens* (4,220) (Lang *et al.*, 2018). Furthermore, the *M. polymorpha* map has an average marker distance of 1.76 cM, which is similar to the genetic map of *P. patens* (1.1 cM).

Nevertheless, our map is based on a significantly fewer set of individuals than in *P. patens*, making recombination rate estimates less precise. The *M. polymorpha* and *P. patens* maps differ in the proportion of nucleotides that could be incorporated into the genetic map. In *P. patens* approximately 99% of the scaffolds were incorporated into the genetic map (Lang *et al.*, 2018), while in *M. polymorpha* we only reached a value of 88%. This difference, however, unlikely represents differences in the completeness of the two genome assemblies for the following reason. In total, approximately 30 Mbp (10 Mbp V and 20 Mbp U) out of the approx. 230 Mbp *M. polymorpha* genome is occupied by the two sex chromosomes (Yamato *et al.*, 2007; Bowman *et al.*, 2017). Because sex chromosomes are non-recombining on most of their lengths, our assembly is expected to consist of the scaffolds anchored to the eight autosomes. Therefore, the total length of our map spanning about 200 Mpb is estimated to cover most of the autosomal complement of the *M. polymorpha* genome (230 Mbp-20 Mbp = 210 Mpb). Altogether, this suggests that our map provides an accurate and close to complete representation of the chromosome-scale structure of the *M. polymorpha* autosomes, and as such can be used to aid map-based cloning approaches or chromosome-scale genomic or epigenomic analyses.

A draft genome assembly is a representative hypothesis of the true nucleotide sequence of a genome. This is especially true for draft genomes primarily relying on relatively short reads like

the *M. polymorpha* genome. During the construction of the linkage map, we found some inconsistencies between the linkage map and the physical position of markers along the assembled scaffolds. These tend to be abundant at the end of the pseudomolecules/linkage groups. Inconsistencies between the linkage and physical maps can be a consequence of errors in the former or the latter, or in both. Therefore, we suggest that physical positions of the pseudomolecules with inconsistencies between the genetic and physical maps needs further attention because most of them may represent misassembled genomic regions.

Telomeric tandem repeat arrays are short

Telomeres of vascular plants are made up of tandem minisatellite repeats and are maintained by the telomerase enzyme (Nelson *et al.*, 2014; Kim & Kim, 2018). Telomeres maintain the integrity and stability of chromosomes. Most vascular plants are characterized by a conserved telomeric repeat unit but telomeric repeats can be variable, be lost or even replaced by transposons. The presence of telomeric repeats were described for both bryophyte genomes (*M. polymorpha* and *P. patens*) and thought to mainly composed of the *Arabidopsis*-type repeats (Suzuki, 2004; Kim & Kim, 2018). Our analyses of the *M. polymorpha* genome suggest that most of the telomeres are either missing from the current assembly or their structure/repeat composition is more variable than previously thought. In particular, we were unable to found specific long tandem repeat arrays in the *M. polymorpha* linkage map typical for the telomeres of vascular plants. This is in line with previous observations that telomeric repeats are less extensive in the *P. patens* genome than in vascular plants (Lang *et al.*, 2018). Nevertheless, the possibility that this is partly a consequence of assembly quality cannot be excluded and remains to be investigated using high-quality assemblies.

Centromeres coincide with high-density of LINE elements but not with typical long satellite repeats

Centromeres in flowering plants are usually consisting of tandem repeat arrays interspersed with some transposable elements (Melters *et al.*, 2013; Oliveira & Torres, 2018). Tandem repeats are highly variable across species concerning their repeat type, repeat unit length and copy number. They may even differ between different chromosomes of the very same species (Birchler *et al.*, 2012). In the bryophyte, *P. patens*, extensive tandem centromeric repeats were not found which may be a consequence of its frequent selfing behavior facilitating the loss of centromeric repeats (Schneider *et al.*, 2016). Alternatively, high copy number tandem repeats may not be characteristic for bryophyte genomes (Lang *et al.*, 2018). Our analyses suggest that similar to *P. patens,* typical centromeric repeats with long repeat unit size and hundreds to thousands of tandem repeat units per array are missing from the *M. polymorpha* genome. The lack of such tandem arrays is unlikely to be primarily due to assembly problems because typical centromeric repeats could be identified using short-read assemblies in other species in a previous study (Melters *et al.*, 2013). Shorter tandem repeat arrays with shorter repeat units are present in the genome but are not specifically restricted to putative centromeric regions but are interspersed. Altogether, our findings imply that the lack of typical long tandem arrays is likely to be a shared feature of bryophyte centromeres, a property that is distinct from most vascular plant genomes analyzed so far (Copenhaver *et al.*, 1999; Birchler *et al.*, 2012; Oliveira & Torres, 2018; Hartley & O'Neill, 2019). The mechanism leading to lower abundance and shorter centromeric repeats in bryophytes compared to the typical centromeric repeats of vascular plants is currently unknown.

Centromeric DNA of land plants is also known to contain transposable elements, mainly retroelements with low abundance (Birchler *et al.*, 2012; Presting, 2018). In contrast, analysis of the *P. patens* genome showed that overall repeat density was unable to demarcate position of centromeres but high density of specific LTR copia elements coincided with the putative position of centromeres (Lang *et al.*, 2018). Our analyses suggest that centromeres of *M. polymorpha* show similar characteristics. More specifically, overall repeat element density did not demarcate the location of putative centromeres in *M. polymorpha*. Nevertheless, we found that maximum density of LINE elements coincided with the centromeres identified based on Marey maps and DNA-methylation status in almost all (six out of the eight) linkage groups. Besides *P. patens*, this is also similar to some algal genomes, where specific LINE elements were proposed to mark centromeric regions (Blanc *et al.*, 2012; Roth *et al.*, 2017). Altogether, our findings suggest that bryophyte centromeres are not characterized by the high abundance of typical centromeric repeat arrays but

rather the accumulation of specific retrotransposons. These elements may accumulate in centromeres because of the preferential occurrence of their insertion sites, alternatively, they may be crucial for centromere function. In either case, the centromere structure revealed is different from that of typically described for vascular plants and seem to be shared by mosses and liverworts and perhaps by the whole group of bryophytes.

The recombinational landscape of M. polymorpha *and* P. patens *is different at a large scale*

The recently published chromosomal-scale assembly of the *P. patens* genome showed that gene density, repeat density, GC content, and recombination were evenly distributed across the chromosomes (Lang *et al.*, 2018). This is in stark contrast to flowering plant genomes, in which all the above mentioned genomic features are usually more abundant in the middle of the chromosome arms and decline towards the centromeres and the telomeres (Heslop-Harrison, 2000; Zhu *et al.*, 2017). Discovery of this special genome structure in the moss *P. patens* raised the question whether this is a feature shared by all bryophyte lineages or whether it is a unique innovation restricted to *P. patens*.

Our analysis suggests that genome structure of *P. patens* and *M. polymorpha* is remarkably similar in some, while radically different in other aspects. On one hand, we found that the spatial distribution of recombination rate variation at the large-scale considerably differs between the *M. polymorpha* and *P. patens* genomes. Recombination rates were greater in the middle of the chromosome arms in *M. polymorpha*, whereas they were more evenly distributed in *P. patens*. Therefore, the recombination landscape of the *M. polymorpha* genome at the large-scale is similar to what is observed in most flowering plants and the evenly distributed recombination rates seem to be a unique feature of *P. patens* (Heslop-Harrison, 2000; Mézard *et al.*, 2007; Haenel *et al.*, 2018). Similarly, we found that GC content was significantly higher in the middle of the chromosome arms in *M. polymorpha*, while it was evenly distributed in the *P. patens* genome. Altogether, these observations suggest that large-scale genome structure of the *M. polymorpha* and *P. patens* genomes is different regarding the spatial distribution of recombination rates and GC content. Whether this difference is due to the contrasting breeding system of the two species, frequent selfing in *P. patens* and obligate outcrossing in *M. polymorpha*, or to species-specific differences in chromatin organization, remains to be determined.

On the other hand, we found that in both species the density of genes and repeat elements did not significantly differ along chromosome arms from centromeres to telomers at a large scale. That is, genes and repeat elements were evenly distributed across chromosome arms in both the *P. patens* and *M. polymorpha* genomes. This is in stark contrast to the observation made in most flowering plant genomes, in which gene density reaches its maximum and repeat density its minimum in the middle of the chromosome arms (Mehrotra & Goyal, 2014). In summary, we observed that the even distribution of repeat and gene features is a shared property of moss and liverwort genomes and is radically different from the spatially clustered distribution of these features in angiosperm genomes. The mechanisms via which these radically different genome architectures in bryophytes and flowering plants are achieved, is currently unknown.

DNA methylation and its effect on recombination rates

DNA methylation is known to be important in modulating recombination rates in the model plant *A. thaliana*, which is realized in a strong anticorrelation between recombination rates and DNA methylation (Yelina *et al.*, 2015; Zhao *et al.*, 2017; Choi *et al.*, 2018; Dluzewska *et al.*, 2018; Tock & Henderson, 2018; Underwood *et al.*, 2018). In line with that, we found that recombination rate and the level of DNA methylation were significantly anticorrelated, both in *M. polymorpha* and *P. patens*. Therefore, our results suggest that the role of DNA methylation in modulating recombination rates are likely to be conserved across bryophytes and flowering plants. This is in spite of the fact that most bryophyte genes have been reported to lack gene body methylation, and the regulatory mechanisms of non-CG methylation are likely divergent between the moss and liverwort lineages (Takuno *et al.*, 2016; Bewick & Schmitz, 2017; Ikeda *et al.*, 2018). We have previously shown, however, that DNA methylation is highly dynamic during the life cycle of *M. polymorpha*, and could detect gene body methylation in the sporophyte, which produces the cells undergoing meiosis (Schmid *et al.*, 2018). It is possible that, as more information on DNA methylation in sporophytes of bryophytes becomes available, potential similarities with angiosperms will become more clear.

Unequal distribution of epigenetic marks, including cytosine methylation, is assumed to primarily contribute to the large-scale variability in recombination rates along the chromosomes of flowering plants (Underwood *et al.*, 2018). In particular, cytosine methylation around the

centromeric regions of *A. thaliana* is about four times higher than in the middle of the chromosome arms (Underwood *et al.*, 2018). Our analysis shows that this is not the case in *M. polymorpha* and *P. patens* (Lang *et al.*, 2018). In particular, methylation levels around or in the putative centromeres are only slightly greater compared to the rest of the chromosomes in *M. polymorpha*. This suggests, that formation, maintenance, and structure of centromeres maybe a potentially unique and shared feature of the two bryophyte genomes, which is radically different from that of flowering plants (see also above).

Gene body methylated genes have similar characteristics as in flowering plants

The functional, evolutionary significance and occurrence of gene body methylation in the lineages of land plants is highly debated (Takuno *et al.*, 2016; Zilberman, 2017; Bewick & Schmitz, 2017; Bewick *et al.*, 2019; Wendte *et al.*, 2019). Some studies reported that gene body methylated genes are present in bryophytes whereas others argued that they are just of artefactual origin (Zemach *et al.*, 2010; Fulneckova *et al.*, 2016; Takuno *et al.*, 2016). Two recent studies on *M. polymorpha* and on the moss *P. patens* found that gene body methylated genes are present in both bryophytes albeit in a low number and/or are restricted to a specific developmental stage (Lang *et al.*, 2018; Schmid *et al.*, 2018). Furthermore, analysis of gene body methylated genes in *P. patens* suggested their potentially divergent function compared to flowering plants. Here we investigated this question further to see whether findings in *P. patens* can be generalized for *M. polymorpha*.

In flowering plants gene body methylated genes are generally more broadly expressed at an intermediate level, are longer, have more exons and have lower GC content than genes with no gene body methylation (Takuno & Gaut, 2012; Bewick *et al.*, 2017; Bewick & Schmitz, 2017; Muyle & Gaut, 2019; Wendte *et al.*, 2019). In contrast, an analysis on *P. patens* genes with gene body methylation showed that they were more specifically expressed, had lower GC content and were less frequently expressed than non-methylated genes (Lang *et al.*, 2018). Our analyses in part contradicts these findings and suggest that characteristics of gene body methylated genes in *M. polymorpha* are more similar to flowering plants than to *P. patens*. In contrast to *P. patens*, gene body methylated genes are longer, have more exons and are more broadly expressed than

nonmethylated genes in *M. polymorpha*. Furthermore, they are expressed at a similar level than non-methylated genes. Nevertheless, gene body methylated genes have lower GC content than those without gene body methylation in both *P. patens* and *M. polymorpha*. Altogether, our data suggest that gene body methylated genes have in part different characteristics in *P. patens* and *M. polymorpha.* We propose that this is not due to a functionally divergent methylation machinery because *M. polymorpha* and *P. patens* share a common set of orthologous genes involved in DNA-methylation (Bowman *et al.*, 2017; Ikeda *et al.*, 2018; Schmid *et al.*, 2018; Aguilar-Cruz *et al.*, 2019). We rather speculate that genomic distribution and characteristics of genes prone to gene body methylation may differ between the two species which may be a consequence of their divergent mating systems and/or large-scale genomic organization (see this publication). Furthermore, we cannot exclude the possibility that the divergent pattern discovered is a consequence of the quality of the DNA-methylation data sets used to define gene body methylated genes in *M. polymorpha* and *P. patens*. The *M. polymorpha* data set describes DNA-methylation status throughout major developmental stages/tissues of the life cycle (Schmid *et al.*, 2018) whereas the *P. patens* data provides DNA-methylation information for a single developmental stage (gametophore) (Lang *et al.*, 2018). If DNA-methylation is as dynamic in *P. patens* as in *M. polymorpha,* the defined set of gene body methylated genes may be highly biased in *P. patens* and can lead to misleading conclusions. We speculate that further data on DNA-methylation in *P. patens* may reveal shared features of gene body methylation with *M. polymorpha* and with flowering plants. More shared features across land plants would provide additional support to the currently proposed model that gene body methylation may be a natural consequence of *CMT3* function in the maintenance of heterochromatin (Wendte *et al.*, 2019). Although bryophytes don't have CMT3-clade CMT genes (Noy *et al.*, 2013), de novo CG and CHH methylation activity of DNMT3 homologs in heterochromatic regions may provide a potential mechanism how gene body methylation may arise in bryophytes (Yaari *et al.*, 2019).

Correlation between genomic features

Recombination rates vary across chromosomes in many organisms and their distribution often correlates with other genomic features. For instance, in most vascular plant species, recombination rates are significantly positively correlated with the density of genes, GC content, and negatively

correlated with the abundance of repetitive elements (Paape *et al.*, 2012; Glémin *et al.*, 2014; Tiley & Burleigh, 2015; Kent *et al.*, 2017). We showed that the patterns of correlation between recombination rate and genomic features mainly followed this general pattern in the *M. polymorpha* genome. We found that recombination rate is positively correlated with the density of predicted genes. We also revealed a negative correlation between recombination rate and the abundance of repeat elements in the genome. These observations are probably due to a greater efficacy of selection in frequently recombining regions of the genome, which allows efficient removal of repeat elements and increases gene density (Charlesworth, 2012; Haenel *et al.*, 2018).

Another common feature of plant genomes is a correlation of recombination rate with GC content. In most plant genomes, recombination rates are positively correlated with GC content but exceptions with negative or no correlation are known, especially in frequent selfers (Marais *et al.*, 2004; Paape *et al.*, 2012; Pessia *et al.*, 2012; Clément *et al.*, 2017). The significant relationship between recombination rate and GC content is thought to be maintained by GC-biased gene conversion after cross-overs (Clément *et al.*, 2017). Alternatively, the correlation could also be maintained by selection for higher GC content or by a recombination-inducing effect of GC-rich genomic regions (Liu *et al.*, 2018). In *M. polymorpha*, we did not find a significant correlation between recombination rate and GC content as it is often seen in flowering plants. Whether this is a result of the resolution of our linkage map or it is a true signal in the *M. polymorpha* genome as was found for duplicated blocks of genes in rice and *Sorghum bicolor*, remains to be investigated (Wang *et al.*, 2009).

Limited collinearity across land plant and bryophyte genomes

Overall, we found very limited collinearity between bryophyte and vascular plant genomes implying that deep divergence since the common ancestor has eroded conserved ancestral gene blocks. This is not unexpected as recent estimates suggest that vascular plants and bryophytes have started to diverge from one another more than 400 million years ago, many million years earlier than the estimated evolutionary origin of the common ancestor of vascular and/or seed plants (Morris *et al.*, 2018b). Similar degradation of collinearity can be seen between monocot and dicot lineages whereas greater collinearity can be observed within dicot and monocot genomes (Murat *et al.*, 2012; Van Bel *et al.*, 2018; Zhao & Schranz, 2019).

A previous study on the *P. patens* genome suggested that regions showing collinearity between the moss and some angiosperms may represent conserved collinear blocks since the most recent common ancestor of land plants (Lang *et al.*, 2018). Our analysis partially contradicts this hypothesis because regions collinear between vascular plants and *M. polymorpha* or *P. patens* turned out to be unique for each of the bryophyte species. Therefore, a more parsimonious explanation of our finding is that *P. patens* and *M. polymorpha* independently retained a different set of collinear regions may be from the common ancestor of land plants. Our gene ontology (GO) analyses of the genes located in these collinear blocks suggest that functional significance may have facilitated retention.

We also found some genomic regions that show collinearity only between *M. polymorpha* and *P. patens*. Recent studies provide evidence that liverworts and mosses form a monophyletic group also likely containing the clade of hornworts (Cox, 2018; Puttick *et al.*, 2018b; de Sousa *et al.*, 2019). This suggests that regions exclusively collinear between the bryophyte species may have been retained since their common ancestor. Our GO analysis suggests that some of these regions potentially host genes with special importance for the overall life cycle of bryophytes.

Intriguingly, we found that collinearity is more restricted between the two bryophyte species than between bryophytes and vascular plants. The common ancestor of mosses and liverworts and/or hornworts is thought to have existed about 10 million years later than the common ancestor of embryophytes (vascular plants and mosses, liverworts, hornworts) (Morris *et al.*, 2018a). Therefore, if the extent of collinearity is proportional to time, we would have expected somewhat more collinearity between the two bryophyte genomes than between bryophytes and vascular plants. Currently, we can only speculate about the processes may lead to less than expected collinearity between the bryophyte genomes. The *M. polymorpha* genome is one of the smallest genomes within liverworts which may arose via secondary reduction and could have contributed to the loss of collinearity (Bainard *et al.*, 2013; Bowman *et al.*, 2017). Furthermore, in contrast to *M. polymorpha*, the *P. patens* genome went through at least two rounds of whole-genome duplications and a rapid proliferation of transposable elements that could have facilitated genome rearrangements (Lang *et al.*, 2018). It is also possible that genome-structure dynamics is running with a faster pace in the two bryophyte lineages than in vascular plants. Finally, we cannot exclude the possibility that our result is an artefact of comparing only one moss and one liverwort genomes. It is possible that including more bryophyte genomes in the analysis we would have

discovered more collinear regions shared by the majority of bryophyte species. Further bryophyte genomes with high-quality genome assemblies are needed to answer this question.

Experimental procedures

Plant material and mapping population

We established a mapping population using two geographically divergent strains, the Japanese male single spore isolate "Tak-1" (Shimamura, 2016) and the Swiss female single spore isolate "IrchF". IrchF was established from a female plant collected at Irchel campus of the University of Zurich, Switzerland, in fall 2012 using surface sterilized gemmae. We achieved induction of sexual reproductive structures and fertilization via established protocols (Ishizaki *et al.*, 2016). Spores were germinated from a single sporophyte on BCD medium (Cove *et al.*, 2009) in a petri dish for four days under continuous light, 22 °C, 300 μEm^2sec^{-1} light intensity, and 60% relative humidity in a growth chamber. From this pool of sporelings, we initially picked 300 single spore isolates using sterile needles and maintained cultures by vegetative propagation on BCD medium.

DNA extraction, ddRAD library preparation, and sequencing

Of the 300 F1 plants, we randomly selected 90 single spore isolates for further genomic analysis. We extracted genomic DNA from these 90 samples and from the parental strains, following a modified CTAB protocol (Rogers & Bendich, 1985). DNA quantity and quality were assessed with a NanoDrop D-1000 (Thermo Fisher Scientific) and by agarose gel electrophoresis. We estimated DNA quantity by a Qubit 1.0 Fluorometer (Invitrogen, Thermo Fisher Scientific).

Subsequent to DNA extraction, we performed double-digest RAD (ddRAD) library preparation for Illumina sequencing, following a modified protocol described in (Baughman *et al.*, 2017). In brief, the genomic DNA was digested by *Eco*RI and *Mse*I, resulting in sticky-end fragments that were labelled by using 92 unique *Eco*RI adaptors containing an in-line barcode. We also extracted DNA from both parental lines (Tak1 and IrchF) and prepared ddRAD libraries following the same procedure. The ddRAD-seq libraries were sequenced on two separate Nextseq500 (Illumina, San Diego, California, USA) flow cells in a single-end mode, producing 125 bp long reads. Each pool

contained libraries of the 90 individuals in equimolar ratios plus libraries of the two parental lines with twice the molarity of the segregants.

Data preprocessing and reference-based haplotyping

Raw data was demultiplexed, and quality filtered with the *process_radtags* program with the "rescue barcodes" (-r), "discard reads with low quality scores" (-q) option implemented in Stacks (v1.46) (Catchen *et al.*, 2013). After this step, we trimmed the reads to exactly 100bp by Trimmomatic (Bolger *et al.*, 2014).

We downloaded the *M. polymorpha* reference draft genome (v3.1) from Phytozomev12 (Goodstein *et al.*, 2012) and aligned each demultiplexed file to the genome with BWA-MEM with default parameters (Li, 2013). The resulting alignments were used to build loci and call haplotypes with pstacks (m = 3 and α = 0.05). After that, we built a reference catalog of RAD loci with cstacks for the two parental lines, using default parameter values, and matched stacks of the progeny against the parental catalog. Finally, we exported genotypes of the parental lines and that of the progeny in the Joinmap 0.4 format, using the genotypes module of Stacks. We discarded markers that occurred in less than 10 individuals of the progeny and did not pass a minimum sequencing depth threshold of five reads. We used the "cross F2" option because both the parental lines and the progeny is haploid with no heterozygosity expected.

Linkage map construction

We constructed a linkage map using R/qtl (Broman *et al.*, 2003) and ASMap (Taylor & Butler, 2017). We excluded individuals with more than 50% missing data and retained only markers that could be genotyped in at least 80% of the individuals. Pairs of genetically highly similar individuals were identified and combined into putative genetic clones (*genClones* command, similarity threshold of 0.95). We excluded 12 markers showing significant segregation distortion (χ^2-test, $p<10^{-8}$, corresponding to a Bonferroni corrected p-value of $p\leq0.05$).

We used the mstmap function of ASMap, with the Kosambi distance function and a p-value threshold of 10^{-10}, to group the markers into linkage groups and to estimate their genetic distance

in cM (centimorgan). To test whether the number of linkage groups is robust against the significance threshold (p-value) applied, we generated linkage groups using multiple p-values ranging from 10^{-5} to 10^{-10}. We re-estimated genetic distances of the markers and ordered them within each linkage group, using the est.map (R/qtl) and the quickEst (ASMap) functions and applying the same thresholds as for the mstmap function. To visually assess the quality of the constructed linkage map, we plotted a heatmap (heatmap, ASMap) of the pairwise recombination rate estimates between markers and the associated LOD (Logarithm of the Odds) scores describing their strengths of linkage. To make our results comparable those of the *P. patens* genome, we obtained segregation data from the original publication (Lang *et al.*, 2018), and reconstructed a linkage map using the very same analytical tools and parameters mentioned above.

Genome anchoring

To anchor our genetic map to the *M. polymorpha* draft genome, we used ALLMAPS (Tang *et al.*, 2015). First, we mapped consensus sequences of the ddRAD tags to the genome draft using blat (Kent, 2002)(blat -fastMap) to ascertain the physical location of all ddRAD-tag markers on the genomic scaffolds of *M. polymorpha*. We kept only markers that showed a perfect match to the genome at least 95% of their lengths and were uniquely mapped. We used this mapping file, the linkage map, and the scaffolds of the draft genome to anchor the draft assembly to the linkage map, employing ALLMAPS' assembly module with the path command. To test the quality of the anchoring process, we calculated non-parametric correlation between physical and genetic distances of mapped markers for each chromosome separately. We generated two genome assembly files. In one file, we stitched together scaffold and contig sequences of the draft assembly with 100 Ns regardless of the size of the gap indicated by the genetic map. In a second file, we used the genetic map to estimate the length of the gaps in base pairs between adjacent scaffolds of the draft assembly. To do this, we first obtained cM/Mb recombination rate estimates by fitting a cubic spline onto the plot of physical and genetic distances for each of the adjacent scaffolds of the gap, and estimated recombination rate as the derivative of the spline. We then used this information to convert the estimated recombination rate between markers spanning the gap into nucleotides. In this assembly, Ns are proportional to the estimated size of the gaps between scaffolds. Finally, we used the liftover tool of ALLMAPS to transfer the genome annotation of the

v3.1 genome (available under https://doi.org/10.5281/zenodo.1117842) to the chromosome-scale assembly.

Marey maps

To estimate local recombination rates along the chromosomes, we used Marey maps (Chakravarti, 1991). This method relies on the comparison of physical and genetic distances along the chromosomes to estimate local recombination rates. Because our genetic map contains between-scaffold gaps, the lengths of which are poorly known, we constructed Marey maps for the map in which gaps are filled in by 100Ns, regardless of the estimated size of the gaps.

To improve the quality of our map files, we filtered them to remove mis-mapped and/or bad markers. We first searched for markers that were incorrectly ordered. We assumed that the reference genome is correctly assembled, and corrected the order and orientation of the genetic map to make it consistent with the assembly. To remove incongruent markers, we searched for the longest common subsequence (LSC) between ranked genomic and physical positions of markers and removed those that were not part of the LSC using slightly modified scripts available in (Corbett-Detig *et al.*, 2015).

Using the genetic and the physical position of markers on the pseudomolecules, we generated Marey maps for both the *M. polymorpha* and the *P. patens* assemblies using the MareyMap R package (Rezvoy *et al.*, 2007; Siberchicot *et al.*, 2017). To estimate local recombination rates (cM/Mb), we employed the loess method that fits a locally adjusted 2nd order polynomial curve to the data points which can adapt to the uneven physical distribution of markers along the chromosomes. We used three window sizes, 0.2, 0.15, and 0.1 (called span) for the local adjustment, that is each window contained 0.2%, 0.15%, and 0.1% of the total number of markers. These values represent a decreasing degree of smoothing with estimates that are more local at smaller span values. The loess method is known to perform well under various degrees of map quality, and provides a good estimation for recombination variation at a large-scale.

We generated Marey maps and recombination rates along pseudomolecules of the *P. patens* v3.3 genome assembly in the very same way. We retrieved physical position of genetic markers on pseudomolecules of the v3.3 assembly from the original publication (Lang *et al.*, 2018). Visual inspection of Marey maps showed either non-monotony or unreliably large recombination rate estimates for chromosomes 6, 13, 25, and 27, which we therefore excluded from any further analyses.

We tested whether there is significant recombination rate variation across the chromosomes via 10 000 bootstrap resampling of the coefficient of variation of the recombination rate estimated (sd/mean * 100). We did this for each of the eight pseudomolecules of *M. polymorpha* and for the 23 chromosomes of *P. patens* (we excluded chromosomes 6, 13, 25, and 27 from the analysis as explained above).

Telomeres and telomeric repeat screening

To identify telomeric repeats we used trf finder (Benson, 1999) and searched for tandem repeats with a unit length between 5 and 15 bp (usual length of telomeric repeats in plants) (Somanathan & Baysdorfer, 2018) using the following options (trf input.fasta 2 7 7 80 10 50 500). We assessed both copy number and localization of the detected tandem repeats to assess their association with telomeric regions. We also mapped the putative "TTTAGGG" *M. polymorpha* telomeric repeat (Suzuki, 2004) to the assembled linkage groups. We then analyzed the spatial distribution of the repeats, assuming that telomeric repeats show a repeat unit size of 5pb-15bp. Because telomeric repeats should represent the longest tandem arrays of such repeats, we plotted the length (the number of repeat units per tandem array) of such tandem arrays along the linkage groups.

Centromeric repeat screening

To search for centromeric repeats, we looked for high copy number tandem repeats with a repeat unit size of between 50-500 bp characteristic of tandem repeats of vascular plant centromeres (Melters *et al.*, 2013)(Copenhaver *et al.*, 1999)(Oliveira & Torres, 2018). To do so, we used trf finder (Benson, 1999) with the following parameters: 2 7 7 80 10 50 500. To assess the position of

putatively centromeric tandem repeats we plotted the frequency of tandem repeat arrays (repeat unit size range 50-500bp) along the eight LGs. Based on the frequency distribution of repeat unit length and the number of repeat units in tandem arrays we used two thresholds: a) repeat unit length ≥10 bps and repeat unit number per array ≥ 10, b) repeat unit length ≥ 30 bps and repeat unit number per array ≥ 10. We defined these two ad hoc thresholds to select tandem repeat arrays made up of long repeat units with large repeat unit number per array relative to their overall genomic distributions. Because these two thresholds were subjectively defined, we also plotted repeat unit length and repeat unit number per array side by side along all LGs including all tandem array with a repeat unit length ≥ 30. Finally, we also searched a candidate centromeric repeat against the assembled genome sequence reported by (Melters *et al.*, 2013) using blastn and an e-value threshold of 10.

Transposable element identification and centromeres

We run Repeatmodeler (Smit & Hubley, 2015) and then Repeatmasker (Smit *et al.*, 2015) with the Viridiplantae repeat data base to search for repeats and classify them according to Repbase (Bao *et al.*, 2015). We used the default values for both Repeatmodeler and Repeatmasker. After that we classified repeat elements into the following categories and plotted their distribution along the eight linkage groups in 500kb wide non-overlapping windows: a.) DNA transposons (TE), b.) Retroelements, c.) LTR retrotransposons, d.) non-LTR retrotransposons, e.) Gypsy (within LTR retrotransposons), f.) Copia (within LTR retrotransposons), g.) SINES (non-LTR retrotransposons), h.) LINES (non-LTR retrotransposons), i.) Unclassified repeats. For each of these, we then visually assessed whether density of elements coincided with the putative centromeric positions we identified based on the Marey map and DNA methylation profiles. We did not include various repeat elements into this analysis (e.g. helitrons etc.) because they were rare or occurred only on a single linkage group.

Recombination rate variation along the chromosomes and its correlation with select genomic features

To statistically test whether recombination rate was evenly or unevenly distributed along the chromosomes, we ran a nested ANOVA model with a main effect of chromosomal position nested within the factor of chromosomes. To do so, we obtained recombination rate estimates in the middle of each 500kb nonoverlapping window using our Marey maps for each three span factors separately. We then divided each chromosome arm (from the centromere to the telomere) into four equal-sized segments (main factor "chromosomal position"), and ran linear models in R (lm()) function) to test the overall significance of recombination rate variation among the four segments. We did this both for the *P. patens* and the *M. polymorpha* genome assembly, and for each of the three recombination rate estimates obtained with the three different span factors separately. To make sure that we have a sufficient number of observations in each quartile, we only included the longer arm of the chromosome in the analyses for highly acrocentric chromosomes. In *P. patens*, we only used the longer arm to carry out this test for chromosomes 23, 22 ,21 ,19 ,18 ,17 ,16 ,15 ,14 ,12 ,10 ,9 , 4, and 3. In *M. polymorpha*, we used only the longer arm in linkage group (LG) 1, which was strongly acrocentric. For all other chromosomes both arms were used. We employed the very same test to investigate whether gene density, repeat density, and GC content in 500kb nonoverlapping windows are evenly distributed across the chromosome arms. Gene density, repeat density, and GC content in genomic windows were obtained as described in the paragraph below.

To investigate the relationship between recombination rate and select genomic features, we correlated (Spearman`s rank-correlation) our recombination rate estimates (Marey map: span 0.1, 0.15 and 0.2) with multiple genomic parameters using a nonoverlapping sliding window size of 500 kb. We calculated gene density (number of genes/window and proportion of nucleotides in genes/window), repeat density (number of repeat features/window or proportion of bases in repeats/window), GC content using our assembly and the bedtools coverage command (Quinlan & Hall, 2010)(https://github.com/arq5x/bedtools2). We then calculated the nonparametric correlation of features with that of the estimated recombination rate for each of the three span value (0.1, 0.15, 0.2) separately.

DNA methylation and recombination rate

For *M. polymorpha*, we used data from (Schmid *et al.*, 2018) to obtain the spatial distribution of methylation along the eight linkage groups in all three sequence contexts. We used custom scripts

to lift over methylation bedgraphs to the linkage map, and calculated the average proportion of methylated cytosines in 500kb non-overlapping windows as described above. Because cross-overs occur during meiosis, we used methylation data from the sporophyte stage to test the correlation between recombination rate and level of methylation. We then calculated Spearman's nonparametric rank correlation between methylation values and recombination rate estimates in 500kb windows for each methylation context and for each of the three span values (recombination rate) separately. We carried out the very same analyses for the *P. patens* genome, for which we retrieved methylation data from (Lang et al., 2018). We excluded chromosomes 6, 13, 25, and 27 from the analysis for reasons explained above. To contrast the spatial distribution of methylation in *M. polymorpha* with that of flowering plants, we used whole-genome bisulfite sequencing data available for *A. thaliana* (Yelina et al. 2015). We mapped the *A. thaliana* data to the reference genome (TAIR, version 10) and obtained methylation calls per cytosine using the very same pipeline described above. We also estimated recombination rates along the five *A. thaliana* chromosomes in the same way described above using the Marey map published in (Wright, Agrawal, & Bureau, 2003).

Gene body methylated genes

We carried out two separate analyses to investigate the number and characteristics of gene body methylated genes in *M. polymorpha*; a.) defining gene body methylation as a discrete character, and b.) using gene body methylation as a continuous character. Genes with at least one methylated CpG position (each site`s methylation level had to pass the 90% methylation level threshold) in their gene body (from start to the end position including introns as defined in the gff file) were treated as being gene body methylated when defining it as a discrete character. This is a similar threshold used in a previous study on *P. patens* genes (Lang *et al.*, 2018). Because *M. polymorpha* methylation varies radically throughout development and tissue types, methylation level of each CpG position was calculated as the maximum percentage value across all tissues/developmental stages investigated previously in (Schmid *et al.*, 2018). As explained above, we also carried out our analyses defining gene body methylation as a continuous trait. To do so, we correlated the number of methylated CpG sites per gene body (definition of methylated sites see above) with various genomic and gene expression variables (see below).

Using gene body methylation as a discrete or continuous variable, we tested whether gene body methylated genes are longer, have more exons, have higher GC content, show less tissues specificity in their expression and whether are more silenced than their non-methylated counterparts. Descriptive statistics of gene features were retrieved from the gff file. We obtained gene expression data from Supplementary material 3 of (Bowman *et al.*, 2017) and calculated Tau, a descriptor of expression specificity, as described in (Kryuchkova-Mostacci & Robinson-Rechavi, 2016). When treating gene body methylation as a binary character, we used Wilcox-tests to compare two medians (methylated vs. non-methylated genes). Non-parametric Spearman rank-correlations were employed when using gene body methylation as a continuous character. Finally, we also investigated whether gene body methylated genes have less expression evidence and/or expression correlates with their methylation level. For this analysis, we calculated average expression of genes using the Supplementary material 3 of (Bowman *et al.*, 2017).

Collinearity analysis

We first created ortho groups using 16 species` proteomes using orthofinder2 v2.3.3 (Emms & Kelly, 2015, 2019). The selected 16 species data set included representatives from each major groups of land plants with high-quality genome assemblies (most in chromosomes) plus the sequence-anchored genome of *M. polymorpha*. For each major clade we included species which have experienced different number of large-scale duplication events throughout their history according to (Qiao *et al.*, 2019). Species did not experience recent whole-genome duplications: *Citrus clementina*, *Theobroma cacao*, *Vitis vinifera*, *Prunus persica*, *Cucumis sativus* and *Amborella trichopoda*. Species experienced multiple rounds of recent whole-genome duplications: *Physcomitrella patens*, *Selaginella moellendorfii*, *Zea mays*, *Oryza sativa* v7_JGI, *Brassica oleracea*, *Arabidopsis thaliana*, *Populus trichocarpa*, *Medicago truncatula*, *Daucus carota*. We added the gff file of our linkage map to this data set. Gff files and proteomes for all the other species were retrieved from phytozome12 (Goodstein *et al.*, 2012).

We then used the gene families obtained to carry out collinearity analysis. We used I-ADHore3 (Proost *et al.*, 2012) to detect highly degenerate collinear blocks expected among bryophytes and vascular plants. We carried out collinear segment detection requesting a minimum of 3, 4 and 5 anchor points within each collinear region (gap_size=30, cluster_gap=35, q_value=0.75,

prob_cutoff=0.01, anchor_points=5, alignment_method=gg2, level_2_only=false). Finally, we conducted gene ontology enrichment analyses using the gene set found in the collinear segments using topGO (Alexa & Rahnenfuhrer, 2019) with a p-value threshold of 0.05. We redundancy filtered and visualized results of the GO-enrichment analyses using Revigo (Supek *et al.*, 2011) by applying a semantic similarity threshold of 0.1 of the SimRel similarity measure.

Acknowledgements

This work was supported by the URPP *Evolution in Action*, grants of the Swiss National Science Foundation (PSZ 160004, 131726; UG 31003A_179553), the EU's Horizon 2020 research and innovation program (PSZ and EC, PlantHUB-No 722338), the NSF (SFM DEB-1541005), the Georges and Antoine Claraz Foundation (AN,AK,MW,ISD,PSZ,OS), the Foundation of German Business (sdw, AN), and the Dept. of Systematic and Evolutionary Botany, University of Zurich (HZS).

Conflict of interests

The authors declare that there is no conflict of interests regarding the publication of this manuscript.

Author Contributions

PSZ designed the research, and SFM contributed to refinement of the design. ISD, OS, MW, AK, AN, GP, RM-W, FB, ZB, ZsH, and CP carried out the experiments. ISD, AG-F, OS, RM-W, and PSZ performed the data analyses. PSZ, OS, ISD, EC, UG, SFM wrote the manuscript, and all authors contributed substantially to revisions.

Accession numbers

Sequencing data are available under the ENA study number PRJEB31477 (flow cell 1: ERR3184756, ERR3184777-ERR3184779; flow cell 2: ERR3184780-ERR3184783). Genome assemblies are available on figshare (https://figshare.com/s/e2f6bc816b78557b93ed).

Supporting information

Fig. S1 Physical and genetic position of genetic markers on the *Marchantia polymorpha* linkage map.

Fig. S2 Distribution of repeat array length (number of repeat units in tandem repeat arrays) along the eight linkage groups of *M. polymorpha*.

Fig. S3 Distribution of tandem array length (number of repeat units per array) and repeat unit length.

Fig. S4 Distribution of putative centromeric repeats ([a] repeat unit length ≥10bps and number of repeat units per tandem array≥10, [b] repeat unit length ≥10bps and number of repeat units per tandem array≥30; see methods) along the eight linkage groups of *M. polymorpha*.

Fig. S5 Recombination rate (cM/Mb) and DNA methylation variation along the eight largest chromosomes of *Physcomitrella patens*.

Fig. S6 Distribution of transposable elements along the eight linkage groups of *M. polymorpha*.

Fig. S7 Recombination rate (cM/Mb) and DNA methylation variation along the five chromosomes of *Arabidopsis thaliana.*

Fig. S8 Distribution of the collinear regions on the chromosomes of *M. polymorpha* and *P. patens.*

Fig. S9 GO enrichment analysis of the collinear regions.

Table S1 Tandem repeat arrays found in the *M. polymorpha* linkage map assembly. Raw output of the tandem repeats finder software (Benson, 1999).

Table S2 Coefficient of variation (CV) of recombination rates in 500kb windows along the eight linkage groups of *M. polymorpha*.

Table S3 Coefficient of variation (CV) of recombination rates in 500kb windows along the chromosomes of *P. patens*.

Table S4 Results of the i-ADHore analysis using a minimum of five anchor points per collinear block.

Table S5 Results of the i-ADHore analysis using a minimum of four anchor points per collinear block.

Table S6 Results of the i-ADHore analysis using a minimum of three anchor points per collinear block.

References

Alexa A, Rahnenfuhrer J. 2019. topGO: Enrichment Analysis for Gene Ontology. R package version 2.36.0.

Aguilar-Cruz A, Grimanelli D, Haseloff J, Arteaga-Vázquez MA. **2019**. DNA methylation in Marchantia polymorpha. *New Phytologist* **223**: 575–581.

Alexa A, Rahnenfuhrer J. **2019**. topGO: Enrichment Analysis for Gene Ontology. : R package version 2.36.0.

Allen CE. **1917**. A chromosome difference correlated with sex differences in Sphærocarpos. *Science* **46**: 466–467.

Allen CE. **1945**. The Genetics of Bryophytes. II. *Botanical Review* **1**: 260–287.

Bainard JD, Forrest LL, Goffinet B, Newmaster SG. **2013**. Nuclear DNA content variation and evolution in liverworts. *Molecular phylogenetics and evolution* **68**: 619–27.

Bao W, Kojima KK, Kohany O. **2015**. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**: 4–9.

Baughman JT, Payton AC, Paasch AE, Fisher KM, McDaniel SF. **2017**. Multiple factors influence population sex ratios in the Mojave Desert moss Syntrichia caninervis. *American Journal of Botany* **104**: 733–742.

Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van De Peer Y, Coppens F, Vandepoele K. **2018**. PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research* **46**: D1190–D1196.

Benson G. **1999**. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**: 573–80.

Bewick AJ, Niederhuth CE, Ji L, Rohr NA, Griffin PT, Leebens-Mack J, Schmitz RJ. **2017**. The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biology* **18**: 1–13.

Bewick AJ, Schmitz RJ. **2017**. Gene body DNA methylation in plants. *Current opinion in plant biology* **36**: 103–110.

Bewick AJ, Zhang Y, Wendte JM, Zhang X, Schmitz RJ. **2019**. Evolutionary and Experimental Loss of Gene Body Methylation and Its Consequence to Gene Expression. *G3: Genes|Genomes|Genetics* **9**: g3.400365.2019.

Birchler JA, Gao Z, Han F. **2012**. Plant centromeres. *Plant Cytogenetics: Genome Structure and Chromosome Function* **9**: 133–142.

Bischler H. **1986**. Marchantia polymorpha l.s. lat. karyotype analysis. *journal of Hattory*

Botanical Laboratory **60**: 105–117.

Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S, *et al.* 2012. The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome biology* **13**: R39.

Bolger AM, Lohse M, Usadel B. **2014**. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bowman JL, Araki T, Kohchi T. **2016**. *Marchantia* : Past, Present and Future. *Plant and Cell Physiology* **57**: pcw023.

Bowman JL, Kohchi T, Yamato KT, Jenkins J, Shu S, Ishizaki K, Yamaoka S, Nishihama R, Nakamura Y, Berger F, *et al.* 2017. Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. *Cell* **171**: 287-304.e15.

Breuninger H, Thamm A, Streubel S, Sakayama H, Nishiyama T, Dolan L. **2016**. Diversification of a Transcription Factor Family Led to the Evolution of Antagonistically Acting Genetic Regulators of Root Hair Growth. *Current Biology* **26**: 1622–1628.

Broman KW, Wu H, Sen Ś, Churchill GA. **2003**. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**: 889–890.

Brzyski JR, Stieha CR, Nicholas McLetchie D. **2018**. The impact of asexual and sexual reproduction in spatial genetic structure within and between populations of the dioecious plant Marchantia inflexa (Marchantiaceae). *Annals of Botany*: 1–11.

Burgeff H. **1943**. *Genetische Studien an Marchantia. Einführung einer neuen Pflanzenfamilie in die genetische Wissenschaft*. Jena: Verlag von Gustav Fischer.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. **2013**. Stacks: An analysis tool set for population genomics. *Molecular Ecology* **22**: 3124–3140.

Chakravarti A. **1991**. A graphical representation of genetic and physical maps: The Marey map. *Genomics* **11**: 219–222.

Charlesworth B. **2012**. The effects of deleterious mutations on evolution at linked sites. *Genetics*

190: 5–22.

Choi K, Zhao X, Tock AJ, Lambing C, Underwood CJ, Hardcastle TJ, Serra H, Kim J, Cho HS, Kim J, *et al.* 2018. Nucleosomes and DNA methylation shape meiotic DSB frequency in Arabidopsis thaliana transposons and gene regulatory regions. *Genome Research* **28**: 532–546.

Clément Y, Sarah G, Holtz Y, Homa F, Pointet S, Contreras S, Nabholz B, Sabot F, Sauné L, Ardisson M, *et al.* 2017. Evolutionary forces affecting synonymous variations in plant genomes. *PLoS Genetics* **13**: 1–28.

Copenhaver GP, Nickel K, Kuromori T, Benito MI, Kaul S, Lin X, Bevan M, Murphy G, Harris B, Parnell LD, *et al.* 1999. Genetic definition and sequence analysis of Arabidopsis centromeres. *Science (New York, N.Y.)* **286**: 2468–74.

Corbett-Detig RB, Hartl DL, Sackton TB. **2015**. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLoS Biology* **13**: 1–25.

Cove DJ, Perroud P-F, Charron AJ, McDaniel SF, Khandelwal A, Quatrano RS. **2009**. Culturing the Moss Physcomitrella patens. *Cold Spring Harbor Protocols* **2009**: pdb.prot5136-pdb.prot5136.

Cox CJ. **2018**. Land Plant Molecular Phylogenetics: A Review with Comments on Evaluating Incongruence Among Phylogenies. *Critical Reviews in Plant Sciences* **0**: 1–15.

Dluzewska J, Szymanska M, Ziolkowski PA. **2018**. Where to Cross Over? Defining Crossover Sites in Plants. *Frontiers in genetics* **9**: 609.

Emms DM, Kelly S. **2015**. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**: 1–14.

Emms DM, Kelly S. **2019**. OrthoFinder: phylogenetic orthology inference for comparative genomics. *bioRxiv* **33**.

Fajkus P, Peška V, Sitová Z, Fulnečková J, Dvořáčková M, Gogela R, Sýkorová E, Hapala J, Fajkus J. **2016**. Allium telomeres unmasked: The unusual telomeric sequence (CTCGGTTATGGG)n is synthesized by telomerase. *Plant Journal* **85**: 337–347.

Fitting H. **1936**. Untersuchungen über die Induktion der Dorsiventralität bei den Marehantieen brutkörpern. II. Die Schwerkraft als Induktor der Dorsiventralität. *Jahrbüch. Wiss. Bot.* **82**: 696–740.

Fulneckova J, Sevcikova T, Lukesova A, Sykorova E. **2016**. Erratum to: Transitions between the Arabidopsis-type and the human-type telomere sequence in green algae (clade Caudivolvoxa, Chlamydomonadales) (Chromosoma, 10.1007/s00412-015-0557-2). *Chromosoma* **125**: 453.

Glémin S, Clément Y, David J, Ressayre A, Glémin S, Clément Y, David J, Ressayre A. **2014**. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends in Genetics* **30**: 263–70.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, *et al.* 2012. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research* **40**: 1178–1186.

Groen KE, Stieha CR, Crowley PH, McLetchie DN. **2010**. Sex-specific plant responses to light intensity and canopy openness: Implications for spatial segregation of the sexes. *Oecologia* **162**: 561–570.

Haenel Q, Laurentino TG, Roesti M, Berner D. **2018**. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Molecular Ecology* **27**: 2477–2497.

Halbsguth W, Kohlenbach H-W. **1953**. Einige versuche über die Wirkung von Heteroauxin suf die Symmetrieentwicklung der Brutkörperkeimlinge von Marchantia polymorpha L. *Planta* **42**: 349–366.

Hartley G, O'Neill R. **2019**. Centromere Repeats: Hidden Gems of the Genome. *Genes* **10**: 223.

Haupt G. **1932**. Beiträge zur Zytologie der Gattung Marchantia (L.) II. *Z. Induk. Abstammungs-Vererbungsl.* **62**: 367–428.

Hedwig J. **1783**. Theoria Generationis et Fructificationis Plantarum Cryptogamicarum Linnaei, mere propriis Observationibus et Experimentis Superstructa.

Heslop-Harrison JS. **2000**. Comparative genome organization in plants: from sequence and

markers to chromatin and chromosomes. *The Plant cell* **12**: 617–36.

Hofmeister W. **1851**. *Vergleichende Untersuchungen der Keimung, Entfaltung und Fruchtbildildiung höherer Kryptogamen (Moose, Farne, Equisetaceen, Rhizocarpeen und Lycopodiaceen) und der Samenbildung der Coniferen.* Leipzig: Hofmeister, W Verlag.

Honkanen S, Thamm A, Arteaga-Vazquez MA, Dolan L. **2018**. Negative regulation of conserved RSL class I bHLH transcription factors evolved independently among land plants. *eLife* **7**.

Ikeda Y, Nishihama R, Yamaoka S, Arteaga-Vazquez MA, Aguilar-Cruz A, Grimanelli D, Pogorelcnik R, Martienssen RA, Yamato KT, Kohchi T, *et al.* 2018. Loss of CG Methylation in Marchantia polymorpha Causes Disorganization of Cell Division and Reveals Unique DNA Methylation Regulatory Mechanisms of Non-CG Methylation. *Plant & cell physiology* **59**: 2421–2431.

Ishizaki K, Nishihama R, Yamato KT, Kohchi T. **2016**. Molecular genetic tools and techniques for marchantia polymorpha research. *Plant and Cell Physiology* **57**: 262–270.

Jamilena M, Mariotti B, Manzano S. **2008**. Plant sex chromosomes: Molecular structure and function. *Cytogenetic and Genome Research* **120**: 255–264.

Kamisugi Y, Von Stackelberg M, Lang D, Care M, Reski R, Rensing SA, Cuming AC. **2008**. A sequence-anchored genetic linkage map for the moss, Physcomitrella patens. *Plant Journal* **56**: 855–866.

Kent WJ. **2002**. BLAT - The BLAST-like alignment tool. *Genome Research* **12**: 656–664.

Kent T V., Uzunović J, Wright SI. **2017**. Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**: 20160458.

Kim MK, Kim WT. **2018**. Telomere Structure, Function, and Maintenance in Plants. *Journal of Plant Biology* **61**: 131–136.

Kryuchkova-Mostacci N, Robinson-Rechavi M. **2016**. A benchmark of gene expression tissue-specificity metrics. *Briefings in bioinformatics*: 027755.

Lang D, Ullrich KK, Murat F, Fuchs J, Jenkins J, Haas FB, Piednoel M, Gundlach H, Van Bel M, Meyberg R, *et al.* 2018. The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant Journal* **93**: 515–533.

Li H. **2013**. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv* **00**: 3.

Liu H, Huang J, Sun X, Li J, Hu Y, Yu L, Liti G, Tian D, Hurst LD, Yang S. **2018**. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nature ecology & evolution* **2**: 164–173.

Marais G, Charlesworth B, Wright SI. **2004**. Recombination and base composition: the case of the highly self-fertilizing plant Arabidopsis thaliana. *Genome biology* **5**: R45.

McDaniel SF, Willis JH, Shaw AJ. **2007**. A linkage map reveals a complex basis for segregation distortion in an interpopulation cross in the moss Ceratodon purpureus. *Genetics* **176**: 2489–2500.

Mehrotra S, Goyal V. **2014**. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. *Genomics, Proteomics and Bioinformatics* **12**: 164–171.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, *et al.* 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome biology* **14**: R10.

Mézard C, Vignard J, Drouaud J, Mercier R. **2007**. The road to crossovers: plants have their say. *Trends in Genetics* **23**: 91–99.

Mirbel C-F. **1835**. Researches anatomiques et physiologiques sur le Marchantia polymorpha, pour servir a l'histoire du tissu cellulaire, de l'épiderme et des stomates. *Me´m. Acad. R. Soc. Inst. France* **13**: 337–436.

Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, Wellman CH, Yang Z, Schneider H, Donoghue PCJ. **2018a**. The timescale of early land plant evolution. *Proc. Natl. Acad. Sci. U. S. A.* **115**: E2274–E2283.

Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, Wellman CH, Yang Z, Schneider H, Donoghue PCJ. **2018b**. The timescale of early land plant evolution. *Proceedings of*

the National Academy of Sciences of the United States of America **115**: E2274–E2283.

Murat F, Van De Peer Y, Salse J. **2012**. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biology and Evolution* **4**: 917–928.

Muyle A, Gaut BS. **2019**. Loss of Gene Body Methylation in Eutrema salsugineum Is Associated with Reduced Gene Expression (M Purugganan, Ed.). *Molecular Biology and Evolution* **36**: 155–158.

Nakayama S, Fujishita M, Fujishita M, Sone T, Ohyama K. **2001**. Additional locus of rDNA sequence specific to the X chromosome of the liverwort, Marchantia polymorpha. *Chromosome Research* **9**: 469–473.

Nambiar M, Smith GR. **2016**. Repression of harmful meiotic recombination in centromeric regions. *Seminars in Cell & Developmental Biology* **54**: 188–197.

Nelson ADL, Beilstein MA, Shippen DE. **2014**. Plant Telomeres and Telomerase. In: Howell SH, ed. Molecular Biology. New York, NY: Springer New York, 25–49.

Noy C, Rafael M, Rachel Y. **2013**. A single CMT methyltransferase homolog is involved in CHG DNA methylation and development of Physcomitrella patens.

Oliveira LC, Torres GA. **2018**. Plant centromeres: genetics, epigenetics and evolution. *Molecular Biology Reports* **45**: 1491–1497.

Ono K. **1976**. Cytological observations on the calluses and the restored thalluses in Marchantia polymorpha. *The Japanese journal of genetics* **51**: 11–18.

Oppenheimer H. **1922**. Das Unterbleiben der Keimung in den Behältern der Mutterpflanze. *Sitzungsber. Kaiserlichen Akad. Wiss.* **131**: 279–312.

Otto K-R. **1976**. Der Einfluß von äußeren Faktoren auf die Bildung von Primärrhizoiden bei Brutkörpern von Marchantia polymorpha L. *Zeitschrift für Pflanzenphysiologie* **80**: 189–196.

Otto K-R, Halbsguth W. **1976**. Die Förderung der Bildung von Primärrhizoiden an Brutkörpern von Marchantia polymorpha L. durch Licht und IES. *Zeitschrift für Pflanzenphysiologie* **80**: 197–205.

Paape T, Zhou P, Branca A, Briskine R, Young N, Tiffin P. **2012**. Fine-scale population recombination rates, hotspots, and correlates of recombination in the Medicago truncatula genome. *Genome Biology and Evolution* **4**: 726–737.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. **2012**. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biology and Evolution* **4**: 675–682.

Presting GG. **2018**. Centromeric retrotransposons and centromere function. *Current Opinion in Genetics and Development* **49**: 79–84.

Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de Peer Y, Vandepoele K. **2012**. i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic acids research* **40**: e11.

Puttick MN, Morris JL, Williams TA, Cox CJ, Edwards D, Kenrick P, Pressel S, Wellman CH, Schneider H, Pisani D, *et al.* 2018a. The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Curr. Biol.* **28**: 733-745.e2.

Puttick MN, Morris JL, Williams TA, Cox CJ, Edwards D, Kenrick P, Pressel S, Wellman CH, Schneider H, Pisani D, *et al.* 2018b. The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Current Biology* **28**: 733-745.e2.

Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. **2019**. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology* **20**: 1–23.

Quinlan AR, Hall IM. **2010**. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Rezvoy C, Charif D, Guéguen L, Marais GAB. **2007**. MareyMap: An R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* **23**: 2188–2189.

Rogers SO, Bendich AJ. **1985**. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Molecular Bioloy*: 69–76.

Roth MS, Cokus SJ, Gallaher SD, Walter A, Lopez D, Erickson E, Endelman B, Westcott D, Larabell CA, Merchant SS, *et al.* 2017. Chromosome-level genome assembly and transcriptome of the green alga Chromochloris zofingiensis illuminates astaxanthin production . *Proceedings of*

the National Academy of Sciences **114**: E4296–E4305.

Schmid MW, Giraldo-Fonseca A, Rövekamp M, Smetanin D, Bowman JL, Grossniklaus U. **2018**. Extensive epigenetic reprogramming during the life cycle of Marchantia polymorpha. *Genome Biology* **19**: 9.

Schmidel CC. **1762**. *Icones Plantarum et analyses partium aeri incisae atque vivis coloribus insgnitae adiectis indicibus nominum necessariis figurarum explicationibus et brevibus animaduersionibus*. Nürnberg: Johann Christoph Keller.

Schneider KL, Xie Z, Wolfgruber TK, Presting GG. **2016**. Inbreeding drives maize centromere evolution. *Proceedings of the National Academy of Sciences* **113**: E987–E996.

Shimamura M. **2016**. Marchantia polymorpha: Taxonomy, phylogeny and morphology of a model system. *Plant and Cell Physiology* **57**: 230–256.

Siberchicot A, Bessy A, Guéguen L, Marais GAB. **2017**. Marey map online: A user-friendly web application and database service for estimating recombination rates using physical and geneticmaps. *Genome Biology and Evolution* **9**: 2506–2509.

Smit A, Hubley R. **2015**. RepeatModeler Open-1.0. : <http://www.repeatmasker.org>.

Smit A, Hubley R, Green P. **2015**. RepeatMasker Open-4.0. : <http://www.repeatmasker.org>.

Somanathan I, Baysdorfer C. **2018**. A bioinformatics approach to identify telomere sequences. *BioTechniques* **65**: 20–25.

de Sousa F, Foster PG, Donoghue PCJ, Schneider H, Cox CJ. **2019**. Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.). *New Phytologist* **222**: 565–575.

Stark LR, McLetchie DN, Mishler BD. **2005**. Sex expression, plant size, and spatial segregation of the sexes across a stress gradient in the desert moss Syntrichia caninervis. *The Bryologist* **108**: 183–193.

Stieha CR, Middleton AR, Stieha JK, Trott SH, Mcletchie DN. **2014**. The dispersal process of asexual propagules and the contribution to population persistence in Marchantia (Marchantiaceae).

American Journal of Botany **101**: 348–356.

Supek F, Bošnjak M, Škunca N, Šmuc T. **2011**. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms (C Gibas, Ed.). *PLoS ONE* **6**: e21800.

Suzuki K. **2004**. Characterization of telomere DNA among five species of pteridophytes and bryophytes. *Journal of Bryology* **26**: 175–180.

Sykorova E, Lim KY, Chase MW, Knapp S, Leitch IJ, Leitch AR, Fajkus J. **2003**. The absence of Arabidopsis-type telomeres in Cestrum and closely related genera Vestia and Sessea (Solanaceae): First evidence from eudicots. *Plant Journal* **34**: 283–291.

Takuno S, Gaut BS. **2012**. Body-Methylated Genes in Arabidopsis thaliana Are Functionally Important and Evolve Slowly. *Molecular Biology and Evolution* **29**: 219–227.

Takuno S, Ran J-H, Gaut BS. **2016**. Evolutionary patterns of genic DNA methylation vary across land plants. *Nature plants* **2**: 15222.

Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J. **2015**. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biology* **16**: 3.

Taylor J, Butler D. **2017**. R Package ASMap: Efficient Genetic Linkage Map Construction and Diagnosis.

Tiley GP, Burleigh G. **2015**. The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. *BMC Evolutionary Biology* **15**: 1–14.

Tock AJ, Henderson IR. **2018**. Hotspots for Initiation of Meiotic Recombination. *Frontiers in Genetics* **9**.

Underwood CJ, Choi K, Lambing C, Zhao X, Serra H, Borges F, Simorowski J, Ernst E, Jacob Y, Henderson IR, *et al.* 2018. Epigenetic activation of meiotic recombination near Arabidopsis thaliana centromeres via loss of H3K9me2 and non-CG DNA methylation. *Genome Research*: 1–13.

Vincenten N, Kuhl LM, Lam I, Oke A, Kerr ARW, Hochwagen A, Fung J, Keeney S, Vader G, Marston AL. **2015**. The kinetochore prevents centromere-proximal crossover recombination

during meiosis. *eLife* **4**: 1–25.

Wang X, Tang H, Bowers JE, Paterson AH. **2009**. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. *Genome Research* **19**: 1026–1032.

Wendte JM, Zhang Y, Ji L, Shi X, Hazarika RR, Shahryary Y, Johannes F, Schmitz RJ. **2019**. Epimutations are associated with CHROMOMETHYLASE 3-induced de novo DNA methylation. *eLife* **8**: 1–27.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, *et al.* 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* **111**: E4859–E4868.

Yaari R, Katz A, Domb K, Harris KD, Zemach A, Ohad N. **2019**. RdDM-independent de novo and heterochromatin DNA methylation by plant CMT and DNMT3 orthologs. *Nature Communications* **10**: 1613.

Yamato KT, Ishizaki K, Fujisawa M, Okada S, Nakayama S, Fujishita M, Bando H, Yodoya K, Hayashi K, Bando T, *et al.* 2007. Gene organization of the liverwort Y chromosome reveals distinct sex chromosome evolution in a haploid system. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 6472–6477.

Yelina NE, Lambing C, Hardcastle TJ, Zhao X, Santos B, Henderson IR. **2015**. DNA methylation epigenetically silences crossover hot spots and controls chromosomal domains of meiotic recombination in Arabidopsis. *Genes Dev* **29**: 2183–2202.

Zemach A, McDaniel IE, Silva P, Zilberman D. **2010**. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**: 916–919.

Zhao X, Bramsiepe J, Van Durme M, Komaki S, Prusicki MA, Maruyama D, Forner J, Medzihradszky A, Wijnker E, Harashima H, *et al.* 2017. RETINOBLASTOMA RELATED1 mediates germline entry in Arabidopsis. *Science* **356**: eaaf6532.

Zhao T, Schranz ME. **2019**. Network-based microsynteny analysis identifies major differences

and genomic outliers in mammalian and angiosperm genomes. *Proceedings of the National Academy of Sciences* **116**: 2165–2174.

Zhu W, Hu B, Becker C, Doğan ES, Berendzen KW, Weigel D, Liu C. **2017**. Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid. *Genome Biology* **18**: 157.

Zilberman D. **2017**. An evolutionary case for functional gene body methylation in plants and animals. *Genome Biology* **18**: 87.

Zimmerman A. **1882**. Über die Einwirkung des Lichtes auf den Marchantienthallus. *Arb. Bot Inst. Wuu¨rzburg* **2**: 665–669.

Table 1 Descriptive statistics of the eight linkage groups (LG1-LG8) recovered for the *M. polymorpha* genome. All Spearman`s Rhos are significant at the p≤0.000001 level.

Linkage group	Number of markers	Length (cM†)	Number of markers per cM	Marker distance	Minimum of marker distance	Maximum of marker distance	Length (in nucleotides when gaps estimated)	Length (in nucleotides, when gaps filled with 100N)	Spearman`s Rho (physical vs. genetic distances)
LG1	598	81.06	0.136	1,663	1,266	5,081	23567953	22641140	0.992
LG2	623	83.60	0.134	1,583	0	3,805	26029825	25782311	0.987
LG3	539	69.76	0.129	1,957	1,266	8,955	26832194	24993663	0.993
LG4	532	102.65	0.193	1,977	1,266	5,081	27599175	25234484	0.996
LG5	454	101.51	0.224	2,256	1,266	8,955	21417235	20768799	0.997
LG6	713	111.49	0.156	1,778	1,266	6,363	29517180	29517180	0.998
LG7	409	76	0.186	1,949	1,266	3,805	23169218	20520495	0.995
LG8	668	86.15	0.129	1,664	1,266	5,081	30690906	28468777	0.997

† Centi Morgan

Table 2 The effect of chromosome and genomic position (quartile) on the density of genomic features in *M. polymorpha* and *P. patens*. We used a nested ANOVA model with two major factors, quartile (four equal-sized chunks of the chromsome arm) and linkage group, quartiles were nested within linkage groups. Statistics for genomic features and recombination rate were estimated in 500kb windows. Signficant p-values are in bold face.

Species		*Marchantia polymorpha*			*Physcomitrella patens*		
Response variable	Factors	df†	F‡	p§	df†	F‡	p§
Recombination rate (span=0.1)							
	Quartile	3	3.9606	**0.0085**	3	0.8854	0.4483

	Linkage group/Chromosome	7	3.3632	**0.0017**	22	0.0773	1.0000
Recombination rate (span=0.15)							
	Quartile	3	5.1830	**0.0016**	3	0.3039	0.8226
	Linkage group/Chromosome	7	2.3762	**0.0220**	22	0.0458	1.0000
Recombination rate (span=0.2)							
	Quartile	3	4.3390	0.1353	3	0.5732	0.6328
	Linkage group/Chromosome	7	1.5963	0.1353	22	0.1336	1.0000
Gene number							
	Quartile	3	2.0173	0.1111	3	3.2896	**0.0203**
	Linkage group/Chromosome	7	0.8575	0.5404	22	1.0192	0.4372
Proportion of bases covered by genes							
	Quartile	3	0.3362	0.7992	3	1.2610	0.2868
	Linkage group/Chromosome	7	5.1952	**1.16E-05**	22	1.1614	0.2760
Proportion of bases in repeats							
	Quartile	3	0.4835	0.6939	3	1.7944	0.1469
	Linkage group/Chromosome	7	1.1124	0.3544	22	1.2563	0.1932

GC content							
	Quartile	3	24.9172	**1.63E-14**	3	1.3452	0.2586
	Linkage group/Chromosome	7	0.1714	0.9908	22	1.0533	0.3950

†degree of freedom

‡F-ratio

§p-value

Table 3 Correlation between genomic features in the *P. patens* and *M. polymorpha* genomes. Values are Spearman's rank correlation coefficients. Correlation coefficients significant at the p≤0.05 level are in bold face. All genomic features were estimated in 500kb windows.

Species	Genomic features	GC content	Number of genes	Proportion of bases covered by genes	Proportion of bases covered by repeats
P. patens					
	Recombination rate span=0.1	**0.2133**	**0.4552**	**0.4535**	**-0.4818**
	Recombination rate span=0.15	**0.2254**	**0.4975**	**0.4951**	**-0.5198**
	Recombination rate span=0.2	**0.2045**	**0.4974**	**0.4873**	**-0.5208**
	GC content		**0.4843**	**0.5214**	**-0.3182**
	Proportion of bases covered by genes				**-0.7671**

M. polymorpha				
Recombination rate span=0.1	0.0670	**0.1293**	**0.1034**	**-0.1079**
Recombination rate span=0.15	0.0641	0.0952	0.0573	**-0.1070**
Recombination rate span=0.2	0.0893	0.0981	0.0668	**-0.1154**
GC content		**0.2625**	0.0971	-0.0347
Proportion of bases covered by genes				**-0.5240**

Table 4 Correlation of percent methylation with recomb rate in 500Kb windows in the eight linkage groups of *M. polymorpha* and the *P. patens* genome. We calculated Spearman`s nonparametric correlation coefficient (Rho) for all three sequence contexts and for recombination rate estimates obtained at three levels of smoothing (see span values 0.1, 0.15 and 0.2). All correlation coefficients are significant at the FDR 0.05 level (Benjamini-Hochberg). Signficant p-values are in bold face.

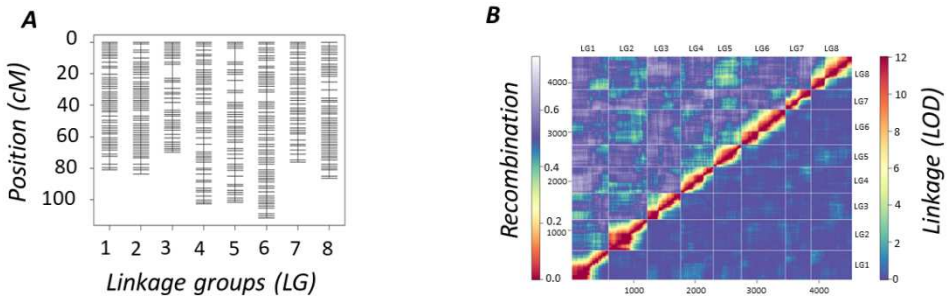
Species	Sequence context	Span		
		0.1	0.15	0.2
M. polymorpha				
	CpG	**-0.1715**	**-0.1643**	**-0.1490**
	CHH	**-0.1176**	**-0.1048**	**-0.1003**
	CpG	**-0.1715**	**-0.1643**	**-0.1490**
P. patens				
	CHG	**-0.5012**	**-0.5444**	**-0.5399**
	CHH	**-0.5146**	**-0.5557**	**-0.5492**
	CpG	**-0.4878**	**-0.5351**	**-0.5356**

Figure legends

Figure 1 Linkage map of *M. polymorpha*. (a) Graphical representation of the linkage map. (b) Plot of estimated recombination rate between genetic marker pairs (upper triangle) vs. LOD score of their linkage (lower triangle) along the eight linkage groups reconstructed.

Figure 2 Recombination rate (cM/Mb) and DNA methylation variation along the eight linkage groups of *M. polymorpha*. Recombination rate and DNA methylation is plotted as dashed and solid lines, respectively. Recombination rates were estimated in 500kb windows with a span value of 0.15. The percentage of methylated cytosines was calculated in three different contexts (CpG in Blue, CHG in green, and CHH in red) and averaged in 500kb windows. Gray areas show putative position of centromeres.

Figure 3 Circos plot of the eight linkage groups of *M. polymorpha*. Density of genomic features were estimated in 500kb windows. *M. polymorpha* chromosomes are assigned to the eight linkage groups. Chromosome outlines are redrawn from (Ono 1976 and Bischler 1986). Large ticks are per megabases (Mbp).



tpj_14602_f1.tif

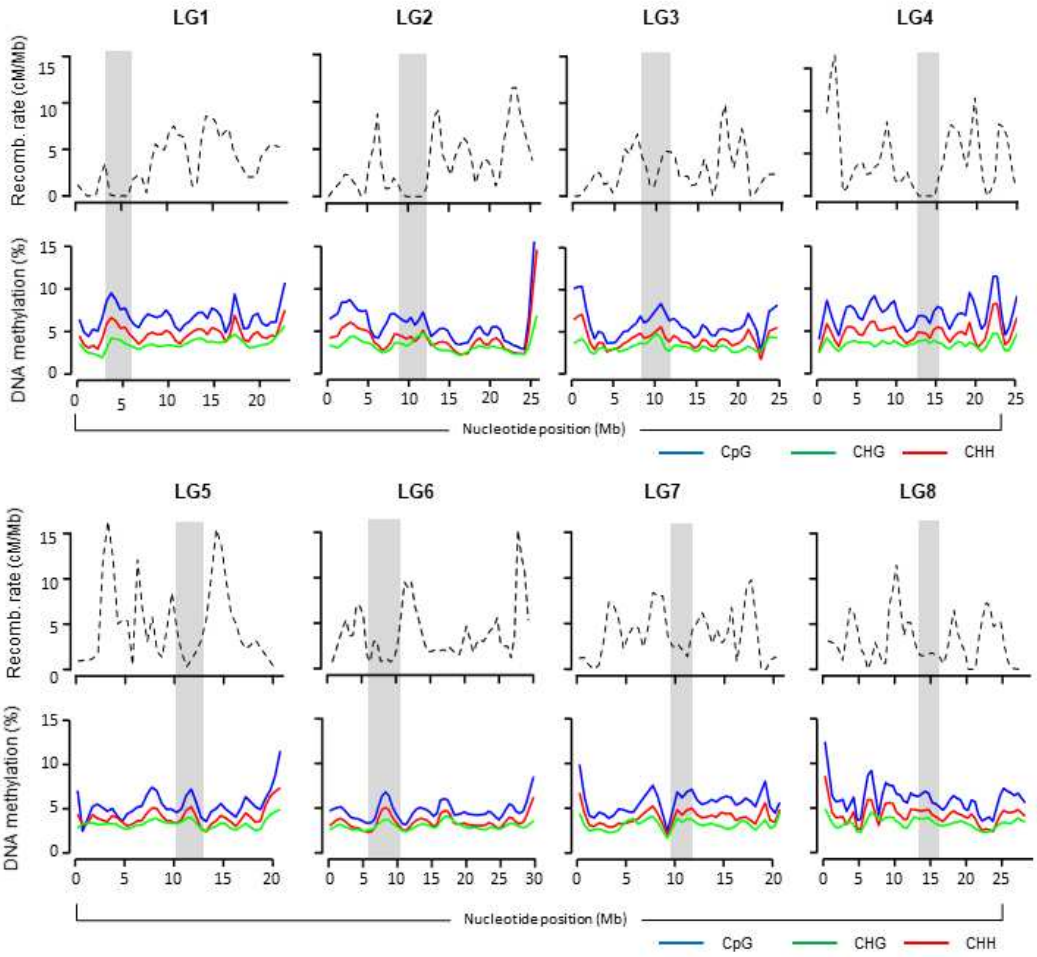
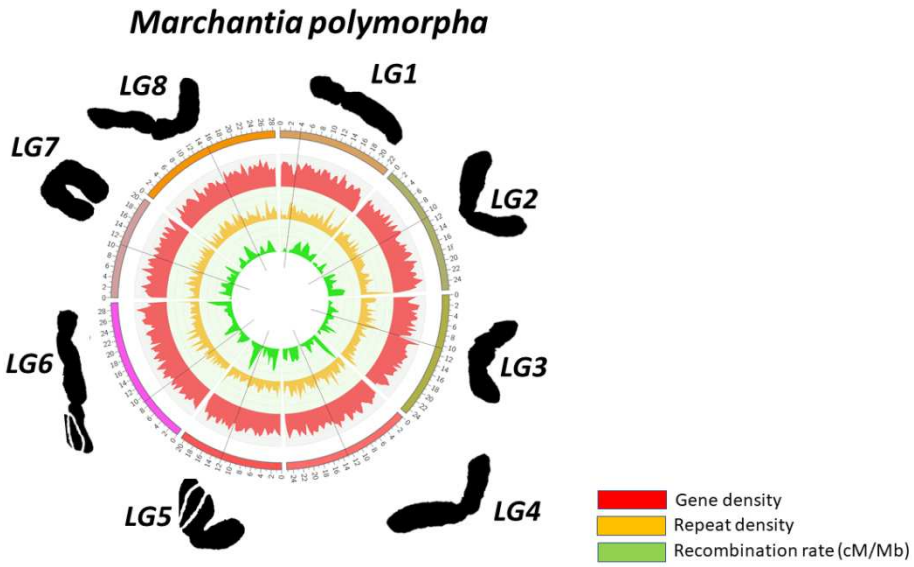


Figure 2 Recombination rate (cM/Mb) and DNA methylation variation along the eight linkage groups of *M. polymorpha*. Recombination rates are more spatially clustered along the linkage groups of *M. polymorpha* than *P. patens* (see Figure S5). Recombination rate and DNA methylation is plotted as dashed and solid lines, respectively. Recombination rates were estimated in 500kb windows with a span value of 0.15. The percentage of methylated cytosines was calculated in three different contexts (CpG in Blue, CHG in green, and CHH in red) and averaged in 500kb windows. Gray areas show putative position of centromeres.

tpj_14602_f2.tif



tpj_14602_f3.tif